Journal of Physics:
Complexity

Celebrating the 5th anniversary of JPhys Complexity

PURPOSE-LED PUBLISHING™

PAPER • **OPEN ACCESS**

Quantifying edge relevance for epidemic spreading via the semi-metric topology of complex networks

To cite this article: David Soriano-Paños *et al* 2025 *J. Phys. Complex.* **6** 035005

View the article online for updates and enhancements.

IOP Publishing J. Phys. Complex. 6 (2025) 035005 https://doi.org/10.1088/2632-072X/adf2ed

Journal of Physics: Complexity

CrossMark

OPEN ACCESS

RECEIVED
20 March 2025

REVISED
7 July 2025

ACCEPTED FOR PUBLICATION
22 July 2025

PUBLISHED
1 August 2025

PAPER

Quantifying edge relevance for epidemic spreading via the semi-metric topology of complex networks

David Soriano-Paños[1,2,*], **Felipe Xavier Costa**[3,4] **and Luis M Rocha**[3,4,*]

1 Departament d'Enginyeria Informàtica i Matemàtiques, Universitat Rovira i Virgili, 43007 Tarragona, Spain
2 GOTHAM lab, Institute for Biocomputation and Physics of Complex Systems, University of Zaragoza, 50018 Zaragoza, Spain
3 Universidade Católica Portuguesa, Católica Medical School, Católica Biomedical Research Centre, Lisbon 1649-023, Portugal
4 School of Systems Science and Industrial Engineering, Binghamton University (State University of New York), Binghamton, NY 13902, United States of America
* Authors to whom any correspondence should be addressed.

E-mail: sorianopanos@gmail.com and rocha@binghamton.edu

Keywords: network sparsification, epidemic dynamics, semi-metric distortion, distance backbone

Supplementary material for this article is available online

Abstract

Sparsification aims at extracting a reduced core of associations that best preserves both the dynamics and topology of networks while reducing the computational cost of simulations. We show that the semi-metric topology of complex networks yields a natural and algebraically-principled sparsification that outperforms existing methods on those goals. Weighted graphs whose edges represent distances between nodes are *semi-metric* when at least one edge breaks the triangle inequality (transitivity). We first confirm with new experiments that the *metric backbone*—a unique subgraph of all edges that obey the triangle inequality and thus preserve all shortest paths—recovers susceptible-infected dynamics over the original non-sparsified graph. This recovery is improved when we remove only those edges that break the triangle inequality significantly, i.e. edges with large semi-metric distortion. Based on these results, we propose the new *semi-metric distortion sparsification* method to progressively sparsify networks in decreasing order of semi-metric distortion. Our method recovers the macro- and micro-level dynamics of epidemic outbreaks better than other methods while also yielding sparser yet connected subgraphs that preserve all shortest paths. Overall, we show that semi-metric distortion overcomes the limitations of edge betweenness in ranking the dynamical relevance of edges not participating in any shortest path, as it quantifies the existence and strength of alternative transmission pathways.

1. Introduction

The advent of network epidemiology in the XXI century [1, 2] has improved our knowledge about how epidemic outbreaks unfold across real interconnected societies. The field's increasing relevance for disease control [3, 4] has been stimulated by the ability to derive realistic network models that characterize human interactions across multiple scales [5, 6]. Indeed, the high resolution characterization of spatiotemporal networks enables actionable interventions to mitigate epidemic outbreaks. For instance, the analysis of mobility networks has shed light on problems as diverse as the risk of importing cases from sources of contagions worldwide [7, 8] and the heterogeneous community transmission observed across a given country [9]. Likewise, including networks to refine epidemic models has enabled their use as reliable benchmarks to assess the short-term impact of mitigation and control policies on spreading dynamics [10].

One of the recurrent problems tackled by network epidemiology is the design of optimal interventions to mitigate an epidemic outbreak while minimally disrupting underlying social, transportation, and trade networks [11, 12]. One such intervention is to focus on nodes, deciding which individuals should be vaccinated first [13–15] or where control resources should be prioritized [16–18] when facing an epidemic. Another family of interventions targets associations between nodes by reshuffling or removing specific edges

[19] to protect the population from the spread of a circulating pathogen. These strategies are usually driven either by edge-centrality measures such as edge betweeness [20–22], which typically characterize only the structure of networks, or may account for dynamical properties of edges [23].

High-resolution network data raises the prospect of digital twins for epidemic forecasting [24]. But to be computationally feasible, *network sparsification* is needed to remove edges that play a minor role in spreading dynamics, without sacrificing important network connectivity and reachability features [24, 25]. Several sparsification methods have been developed over the last couple of decades [26, 27], based on measures of edge importance that focus on different network properties: local node connectivity [28], statistical impact of edge removal [29–33], global structure of paths [34], or redundancy for shortest-path computation in both undirected [35] and directed networks [36]. Other methods to reduce the network size use effective renormalization groups [37], based on either embedding networks on hyperbolic spaces [38] or their Laplacian graph properties [39].

The interplay between network sparsification and spreading dynamics has received much less attention than has the design of targeted strategies to control an outbreak. However, mounting evidence in the literature suggests that network sparsification which relies on global information provides better recovery of spreading dynamics than just removing the weakest connections. For instance, sparsifying a network according to the distribution of effective resistance values (which account for the relevance of a given edge within the ensemble of paths that connect its two end nodes) outperforms weights thresholding in preserving both susceptible-infected (SI) [40] and SI-recovered (SIR) [41] dynamics. Similarly, the study of spreading phenomena through shortest paths in a network has been used to address different problems such as the inference of the source of an outbreak [42] and the expected distribution for the arrival times of the pathogen to different locations [43]. Also focusing on shortest paths, a recently published paper by Correia *et al* [44] shows that the *metric backbone*—a unique and algebraically-principled subgraph composed of the (weighted) edges that obey the triangle inequality and thus compose all shortest paths [35] —provides a more solid foundation for network sparsification in epidemic spread than does relying on local information.

The *metric backbone* is a feature of the *semi-metric topology of complex networks* [35]. More specifically, weighted graphs can always be isomorphically transformed so that their weights represent distances, allowing analysis via the rich mathematics of (probabilistic) metric spaces [45]. This has revealed that the topology of most complex networks derived from real-World data is semi-metric, i.e. the distance between nodes often breaks the triangle inequality [45, 46], resulting in much redundancy for computing shortest paths, since only edges that obey the triangle inequality participate in those [35]. However, determining which features of a given network limit the reconstruction of epidemic outbreaks from the *metric backbone* subgraph remains an open problem, despite previous studies [44]. We tackle this challenge and reveal that improving the reconstruction of spreading dynamics hinges on properties of *semi-metric* edges [45, 46]. These edges are not included in the *metric backbone* and, therefore, do not appear in any shortest path in the network. Specifically, we show that the quality of a reconstruction depends on the interplay between the proportion of semi-metric edges—which is complementary to the relative size of the *metric backbone*— and their associated *semi-metric distortion*, s^m [35, 46], a quantification of how much edges break the triangle inequality. We show that both these measures characterize how the semi-metric topology of complex networks affects spreading dynamics.

Based on these results, we propose and test a new *semi-metric distortion sparsification* (SMDS) method, whereby edges are progressively removed in decreasing order of semi-metric distortion. In other words, SMDS removes first those edges whose (direct) distance weight is much larger than the length of the shortest (indirect) path between the endpoints they connect, thus breaking the triangular inequality the most [35, 45]. SMDS progressively dismantles a network via the (strict, total) ordering of semi-metric distortion values, until the *metric backbone* subgraph is reached. This algebraically-principled sparsification limit ensures the method preserves all shortest paths of the original network—not only the distribution of shortest path length (reachability) [35]. Indeed, sparsification methods that do not preserve the *metric backbone* necessarily alter the original shortest paths. We show that SMDS outperforms weight and effective resistance thresholding in recovering SI, SI-susceptible (SIS), and SIR spreading dynamics, for both synthetic and real networks. Furthermore, it typically results in sparser subgraphs than the other methods.

More generally, the semi-metric topology of networks, which is induced by distance functions that break the triangle inequality, not only defines the edges that are required for shortest paths (the backbone), but also offers a natural and algebraically-principled ranking of the (semi-metric) edges that do not contribute to any shortest path. Even though all semi-metric edges have null edge-betweenness, our results demonstrate that their dynamical importance varies widely and is inversely correlated with semi-metric distortion. While this shows that spreading dynamics does not depend only on shortest paths, it also shows that the edges that are furthest from contributing to any shortest path (those that most break the triangle inequality) are dynamically redundant.

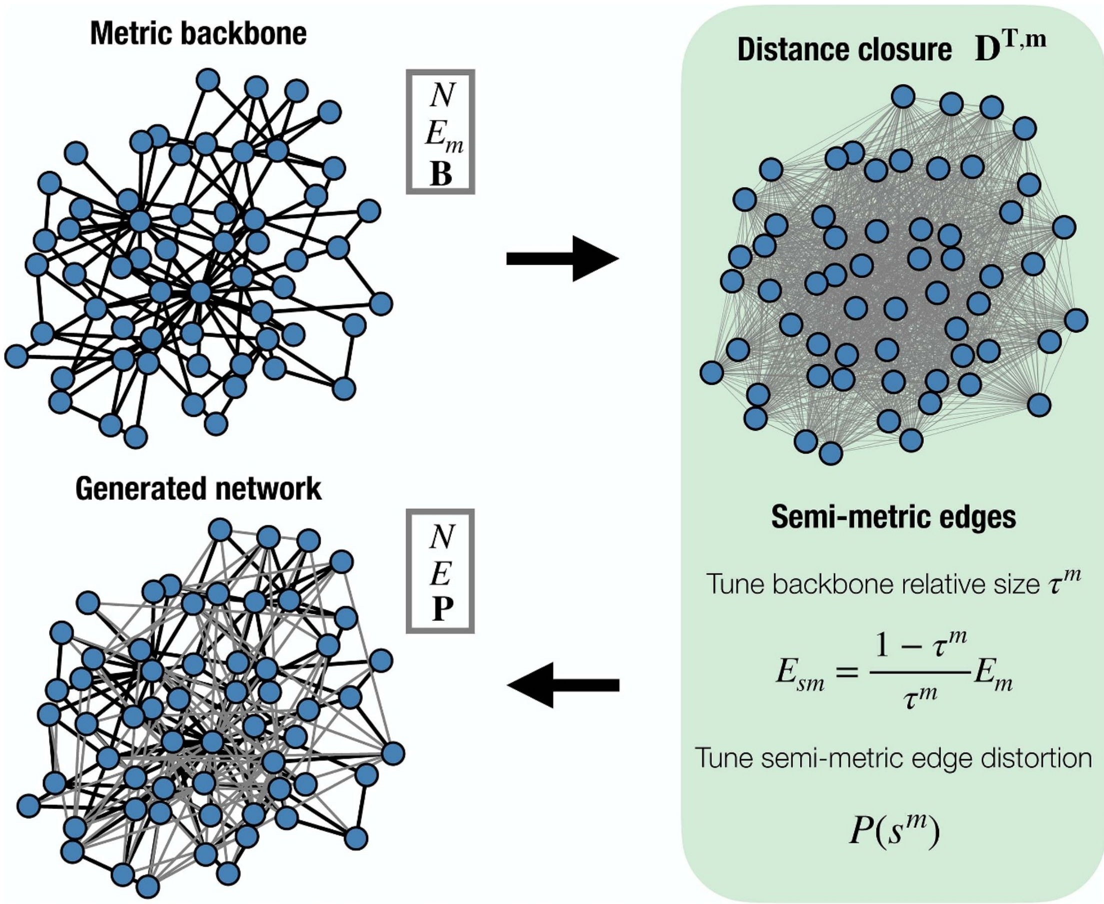

Figure 1. Schematic representation of the construction of synthetic networks with tunable relative size of the backbone τ^m and semi-metric distortion distribution $P(s^m)$. The first step consists of setting the (connected) *metric backbone* of a synthetic network. This *metric backbone* is fully characterized by the number of nodes N, the number of edges E_m and the proximity matrix **B** containing the weights of associations. The next step involves computing the distance closure matrix $\mathbf{D}^{\mathrm{T,m}}$, a graph of the same node set N, whose edge weights denote the shortest distance (length of shortest path) between every pair of nodes. Finally, we generate a synthetic network by adding E_{sm} edges, whose value is computed according to equation (7) in the Methods section, to obtain a desired relative backbone size τ^m. The semi-metric distortion values of the added edges s_{ij}^m are drawn from the target distribution $P(s^m)$. Finally, the resulting proximity matrix **P** is obtained by applying equations (3) and (6) in Methods section to each s_{ij}^m value.

2. Results

2.1. Interplay between SI dynamics and the metric backbone

We first explore different network features that potentially shape the reconstruction of epidemic outbreaks from the metric backbone. To address this challenge, we propose a new method to construct synthetic networks where we can tune the relative size of the *metric backbone* τ^m and the distribution of semi-metric distortion values $P(s^m)$, as shown in figure 1. We refer the reader to the Methods subsection 'Metric backbone and semi-metric distortion' for a complete explanation of the theoretical foundations of the *metric backbone* and the associated semi-metric distortion parameter of an edge. Starting from a *metric backbone* of N nodes and E_m edges, our method adds semi-metric edges to build a network of N nodes and E edges, whose strength of associations are quantified by the proximity matrix **P**. A detailed description of our method to construct synthetic networks is provided in the Methods subsection 'Construction of synthetic networks'.

To focus on the role of the aforementioned features, τ^m and $P(s^m)$, while preserving other empirically relevant structures, the synthetic networks are built from the backbone of a real network. Specifically, we utilize the backbone of a contact network between elementary school students (kindergarten to sixth grade) in Utah (USA) [44, 47]. This *metric backbone* is composed of $N = 339$ nodes and $E_m = 1128$ metric edges, with a heterogeneous distribution of the proximity values (available in figure S1 in the supplementary text). Moreover, for the semi-metric edges, their semi-metric distortion values are drawn from a log-normal distribution via $s^m = 1 + r$ where $r \sim Lognormal(\mu, \sigma)$. Note that this probability function is widespread across different empirical networks [35, 44]. Throughout the manuscript, we fix $\sigma = 1$ and modify the μ value to study the effect of semi-metric edges' relevance with respect to those included in the *metric backbone*.

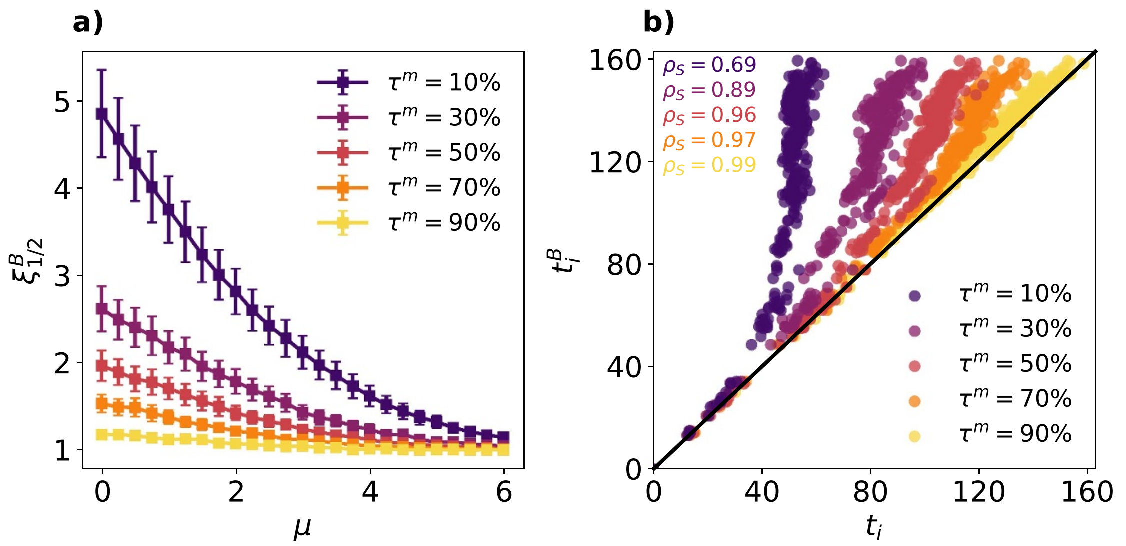

Figure 2. Panel (a): ratio between the time for the disease to reach half the population in the *metric backbone* and in the constructed network $\xi^B_{1/2}$ as a function of the parameter μ, i.e. the logarithm of the median of the semi-metric distortion distribution, and the relative size of the backbone τ^m (color code). For a given seed, we compute the time to reach half the population $t_{1/2}$ as the median value observed across 200 realizations of the SI dynamics. Dots show the average and error bars represent the standard deviation of the ratios obtained for 50 different infectious seeds. Panel (b): average time of infection of each individual i in the *metric backbone* t_i^B as a function of its average time of infection in the entire network t_i for different values of the relative size of the backbone τ^m (color code). These results are obtained by averaging 200 realizations starting from the same initial seed. In both panels, we set $\beta = 0.5$ for the SI dynamics, fix $\sigma = 1$ in the semi-metric distortion distribution and consider the *metric backbone* **B** of a network capturing face-to-face interactions in a elementary school in the US [47].

To assess the extent to which the *metric backbone* can reconstruct epidemic outbreaks, we introduce a single infectious seed, i.e. a single individual initially infected, and run a SI dynamics, which constitutes the simplest framework to simulate spreading phenomena in networks. As detailed in the Methods subsection 'SI model', epidemic outbreaks in the SI model are fully described by the distribution of times of infection for the different nodes across the network. Based on this fact, we assume that the characteristic time scale of SI dynamics in a given network is the time at which half of the population is reached by the outbreak, denoted in here by $t_{1/2}$. Therefore, our research question narrows down to determine whether the characteristic time scale of spreading phenomena in the *metric backbone* **B**, denoted by $t^B_{1/2}$, captures the corresponding measure in the entire network. Note that, in the following comparative measures, the superscript refers to the sparsification method while the subscript to the time of comparison. We define the ratio comparing the times at which half of the population is reached in the backbone and the original network, $\xi^B_{1/2}$, as:

$$\xi^B_{1/2} = \frac{t^B_{1/2}}{t_{1/2}} \,. \tag{1}$$

In absence of stochastic fluctuations, the aforementioned ratio fulfills $\xi^B_{1/2} \geqslant 1$, as the *metric backbone* always removes potential transmission pathways for the virus existing in the original network. In terms of performance, the closer this ratio gets to $\xi^B_{1/2} = 1$, the more faithful the information provided by the *metric backbone* is about the dynamics in the entire network.

Figure 2(a) represents the ratio $\xi^B_{1/2}$ as a function of the semi-metric edges relevance, governed by μ, and the relative size of the backbone τ^m. For large μ values, semi-metric edges are highly dynamically redundant in comparison with the metric ones as $\xi^B_{1/2} \simeq 1$ regardless of the backbone relative size. Therefore, their removal hardly has any influence on the spreading. As edges distances become closer to the shortest path lengths, we observe a critical value μ_c for each τ^m value below which the spreading dynamics gets slower in the *metric backbone* ($\xi^B_{1/2} > 1$). Interestingly, this critical value μ_c increases as the size of the backbone decreases. The latter results imply that the performance of the *metric backbone* is determined by the interplay between both the potential transmission pathways pruned during the sparsification process (governed by τ^m) and their relevance with respect to those kept in the *metric backbone* (the semi-metric distortion, governed by μ).

The previous results show that considering the *metric backbone* as the underlying contact structure might induce global delays in the spreading dynamics. Nonetheless, even in those scenarios, the information obtained from this subgraph can be relevant for disease control if the *metric backbone* allows us to faithfully

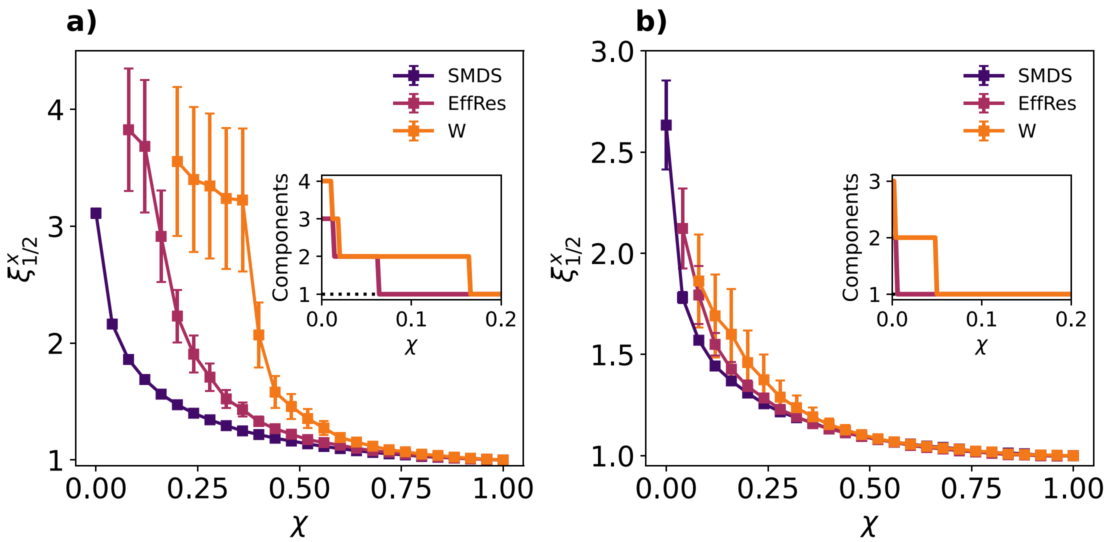

Figure 3. Ratio between the time for the disease to reach half the population in the sparsified network and in the original synthetic network $\xi^x_{1/2}$ as function of the parameter governing the size of the sparsified network χ for each sparsification method x (color code). χ ranges from $\chi = 1$, corresponding to not removing any edge, to $\chi = 0$ where E_{sm} edges are removed. Three different sparsification methods are compared: targeting edges with highest semi-metric distortion values (SMDS), lowest effective resistance values (EffRes) or lowest proximity values (W). The details of the simulations to obtain the ratios are the ones described for figure 2(a). EffRes and W curves are interrupted when breaking the largest connected component of the network. Insets: Number of components in the network as a function of the edges removed χ and the sparsification method (color code). The horizontal dashed line represents the SMDS, which always preserves the largest connected component. The original synthetic networks are constructed considering $\tau^m = 0.1$ and $\mu = 1$ (panel (a)) or $\mu = 2$ (panel (b)) respectively.

rank the different nodes according to their expected time of infection. To check that, we randomly place a single infectious seed in the network and study how the distribution of the individual infection times varies as we alter the properties of the metric backbone. In particular, we fix $\mu = 2$ and explore the role of the relative size of the backbone τ^m in the microscopic reconstruction of outbreaks. Figure 2(b) shows that, even when the *metric backbone* represents just 10% of the edges of the network, it qualitatively captures the epidemic trajectory across the population, as shown by the high Spearman correlation between the distributions obtained for both the network and its backbone ($\rho_S = 0.69$, $p < 0.001$). As expected, the microscopic recovery of the epidemic trajectory is also enhanced as τ^m increases and more transmission pathways are captured in the metric backbone.

2.2. SMDS in synthetic networks

Our findings indicate that semi-metric edges with large s^m values are dynamically redundant, as their removal has negligible effects on SI dynamics. Motivated by this result, here we propose the *SMDS* method, where we sort the edges (i,j) according to their associated distortion s^m_{ij} and progressively remove those with highest values until matching the desired size of the sparsified configuration. We compare the method proposed here with two other sparsification schemes relying on different edges properties: *weights thresholding* and *effective resistance thresholding*. On the one hand, weights thresholding relies on local information, aiming at removing the weakest connections through which transmission of the virus is very unlikely. On the other hand, effective resistance thresholding penalizes path redundancy, as small effective resistance values identify those direct edges connecting nodes which can also exchange information through many other indirect paths. More details on the computation of the effective resistance associated to each edge can be found in the Methods subsection 'Effective resistance'.

To assess and compare performance, we compute the ratio $\xi^x_{1/2}(\chi) = t^x_{1/2}(\chi)/t_{1/2}$ for each subgraph obtained after removing edges according to each sparsification method x. For the SMDS, χ denotes the fraction of semi-metric edges included in the graph, i.e. $E(\chi) = E_m + \chi E_{sm}$. Therefore, $\chi = 1$ corresponds to the entire graph whereas $\chi = 0$ corresponds to the *metric backbone* subgraph. For the other two methods, the sparsified networks comprise $E(\chi)$ edges which are chosen by thresholding proximity weights p_{ij} or effective resistance values p^R_{ij}, defined according to equations (9) and (11) in the Methods section, respectively.

Figure 3(a) shows the comparison of the three sparsification methods in a synthetic network with $\tau^m = 0.1$ and $\mu = 1$. We observe that SMDS outperforms the other two methods in both preserving SI dynamics and keeping the connectedness of the network. Specifically, $\xi_{1/2}$ for SMDS remains closer to 1, showing that edge relevance for SI dynamics is more correlated with their semi-metric distortion than their proximity weights or effective resistance. This demonstrates that even semi-metric edges that do not contribute to any shortest path (they have zero betweenness), are more relevant to SI dynamics the less they break the triangle inequality. In other words, those edges that are nearer to being necessary for computing shortest paths (low semi-metric distortion) play a much more important role in SI dynamics than the edges that are far from contributing to shortest paths (large semi-metric distortion).

It is also important to notice that SMDS does not target any metric edge by definition and therefore guarantees the preservation of the shortest path connecting every pair of nodes, preserving both connectivity and shortest distances. In contrast, the other two methods eventually dismantle the largest connected component, breaking the network into different subgraphs. Note that this phenomenon is more pronounced for thresholding by weights than by effective resistance as the latter harnesses global information and makes the isolation of individual nodes rarer.

Interestingly, increasing μ in the synthetic network reduces the differences between the three methods, as shown in figure 3(b) for networks constructed with $\mu = 2$. In this case, the large distortion of the edges in the synthetic networks turns their weights considerably smaller (figure S1), which reduces the chances that the other thresholding methods target edges in the *metric backbone* when sparsifying the network, as demonstrated in figure S2 in the supplementary text. This supports the role of the *metric backbone* as a primary subgraph sustaining the spread of diseases [44], but it also newly reveals that edges with larger semi-metric distortion rarely contribute to spreading dynamics (notice difference in scale for $\xi^x_{1/2}$ between panels (a) and (b) in figure 3) .

2.3. SMDS in empirical networks

We also study 16 networks built from biological, social and transportation data, as detailed in the Methods subsection 'Empirical Networks'. Despite the small size of this dataset, it contains networks with striking differences in terms of size, weights distributions or degree distributions, as shown in table S1, figures S3 and S4. Therefore, this dataset allows us to explore the generality of the findings made in the synthetic networks and test the performance of SMDS against other sparsification methods in networks of very diverse nature. Figure 4(a) depicts a comparison between the three sparsification methods in regards to recovering SI dynamics for the case the US elementary school social contact network—specifically, time to infect half of the population. It is clear that SMDS presents the same advantages observed in the case of synthetic networks, i.e. better recovery of SI dynamics overall and no disruption of the largest connected component of the network for greater sparsification. A similar analysis for all other networks is provided in figure S5 of the supplementary text, leading to the same overall results.

To quantify the superior recovery of SI dynamics observed for SMDS across empirical networks, we compute the relative differences between the mean ratios observed for the SMDS and each other method x, as a function of the size of the sparsified subgraph χ:

$$\tilde{\Delta\xi}^x_{1/2}(\chi) = \frac{\xi^x_{1/2}(\chi) - \xi^{\mathrm{SMDS}}_{1/2}(\chi)}{\xi^{\mathrm{SMDS}}_{1/2}(\chi)}\,, \tag{2}$$

where, similarly to the synthetic network study (section 2.2), x stands for thresholding by either proximity weights or effective resistance values. Therefore, when $\tilde{\Delta\xi}^x_{1/2} > 0$, SMDS outperforms method x.

Similarly to what was observed in the synthetic network study, figure 4(a) shows little-to-no differences across sparsification methods when removing few edges from the network, as $\xi^x_{1/2} \simeq 1$ when $\chi \simeq 1$ regardless of the method considered. Nonetheless, as more edges get pruned, SMDS typically outperforms the other two methods as shown by the positive median of the distribution of $\tilde{\Delta\xi}^x_{1/2}$ values obtained for the different networks in the dataset (figure 4(b)). However, there are some networks for which $\tilde{\Delta\xi}^x_{1/2} < 0$, which means that the other methods can outperform SMDS in preserving SI dynamics for some empirical networks, such as the Human Connectome network (see figure S5 for the comparison of the methods for each network).

To further support our findings, figure S6 in the supplementary text presents a comparison between the three methods for different stages of an epidemic outbreak. There we observe that SMDS generally allows for a better construction of the dynamics except for early stages of the epidemic spreading, when only 10% of the population has been infected. In that case, the set of infected individuals reflects the composition of very localized infection around the seed node, thus making the global shortest-path information included by semi-metric distortion values less relevant.

Apart from recovering spreading dynamics, a sparsified network should preserve the connectivity of the original network to ensure its functioning. For each sparsification method x, we represent in figure 4(c) the fraction f^x of empirical networks that become disconnected as a function of the size of their sparsified subgraphs. We observe that the empirical networks are quite vulnerable to thresholding by proximity weights, as most of the networks in our dataset eventually break into different components before reaching

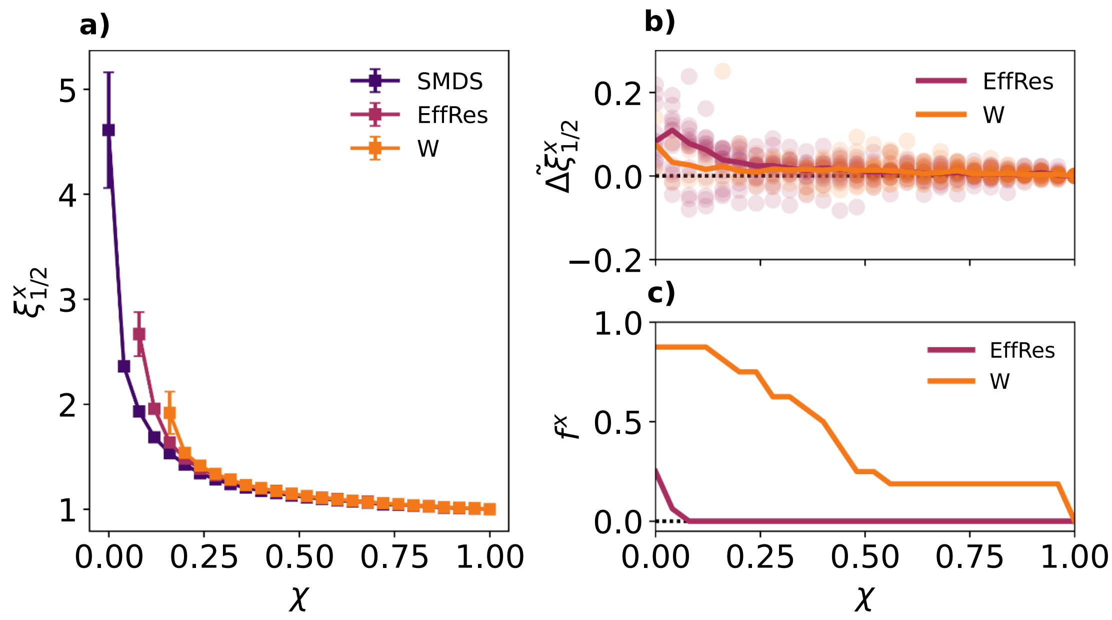

Figure 4. Panel (a): ratio between the time for the disease to reach half the population in the sparsified network and in the original synthetic network $\xi^x_{1/2}$ as a function of the parameter χ, governing the edges removed from the network, and the sparsification method x (color code). The contact network corresponds to face-to-face interactions recorded in an elementary school students (kindergarten to sixth grade) in Utah (USA) [47]. The details of the simulations to obtain the ratios are the ones described for figure 2(a). Panel (b): distribution of the relative difference $\Delta\tilde{\xi}^x_{1/2}$ between the mean ratio associated to each sparsification method x (color code) and SMDS as a function of the size of the sparsified network χ, for all emprical networks. Positive $\Delta\tilde{\xi}^x_{1/2}$ values reveal a better performance of SMDS with respect to the sparsification method x. Dots represent each of the 16 networks included in the dataset under study whereas solid lines show the median $\Delta\tilde{\xi}^x_{1/2}$ value observed across networks for each sparsification method x. Panel (c): fraction of disconnected networks in the dataset of real networks f^x as a function of the size of the sparsified network χ and the sparsification method x (color code). In panels (b) and (c), the horizontal dashed line represents the values obtained for the SMDS.

the size of the *metric backbone* subgraph. While thresholding by effective resistance is more resilient (likely due to the global information encoded in the effective resistance values), 4 out of 16 networks end up breaking into separate components when removing the same number of (but not the same) edges that get removed to obtain the *metric backbone* ($\chi = 0$). Figure S8 in the supplementary text further shows the number of components as a function of the size χ for each of the different networks studied.

Another variable of interest when choosing a suitable sparsification method is the computational time needed to sparsify the network. Table S2 contains the times needed to compute SMDS and effective resistance thresholding in each empirical network showing how the former is typically longer than the latter due to its higher algorithmic complexity. More details on the comparison of algorithmic complexities can be found in the Methods section. Nevertheless, table S2 also shows how SMDS time is negligible in comparison with the gain in computational time obtained from running the simulations in the sparsified configurations. Therefore, SMDS allows preserving SI dynamics and network connectedness while reducing considerably the time needed to analyze epidemic trajectories on networks.

To complete our analysis, we further consider how different sparsification methods preserve epidemic outbreaks under SIS and SIR dynamics, which are more complex compartmental models. These results are shown in figures S9–S12. In both cases, we set the epidemiological parameters to ensure the existence of a widespread epidemic state in all empirical networks. Figure S9 shows that SMDS and effective resistance thresholding lead to similar macroscopic SIS dynamics performance, reaching a similar global epidemic state for the same amount of sparsification. However, in a majority of networks and across sparsification levels, SMDS reaches a fraction of infected individuals at steady state closer to that of the original network, without ever disconnecting it (which effective resistance can do, as shown in figures S5 and S6). Weight thresholding frequently breaks the original networks into components even for modest amounts of sparsification. Though, when this sparsification does not break a network, it preserves slightly better the global epidemic state of the original network. When it comes to microscopic SIS dynamics of an epidemic outbreak, SMDS outperforms effective resistance thresholding in preserving the ranking of nodes according to their individual probabilities of infection (figure S10). Similar results are observed for macroscopic and microscopic SIR dynamics in figures S11 and S12, respectively. These results, along with those reported for the SI dynamics, demonstrate that SMDS is an excellent sparsification method to generate subgraphs preserving spreading dynamics and the connectedness of complex networks.

3. Discussion

Sophisticated data gathering techniques now enable the generation of large-scale networks that represent the architecture of relationships spanning multiple scales in nature, ranging from biochemical interactions to social contacts and transportation flows, all of which greatly contribute to the spread of disease [48, 49]. Despite much scientific progress in understanding how epidemics spread via such networks, the resulting models usually include redundant associations that do not contribute to their dynamics [41, 44, 50, 51]. These redundant associations not only enlarge the networks and their computational analysis [35, 45], but also may obfuscate their topology and causal interaction structure leading to erroneous predictions [52–54]. Network sparsification methods are thus essential to leverage the high spatiotemporal resolution of current datasets toward effective and actionable models such as digital twins [24].

Here we investigate whether the semi-metric topology of weighted graphs provides reliable guidance for sparsifying a network while preserving spreading dynamics. To do so, first we study the reconstruction of SI outbreaks from the metric backbone, which is composed of the subset of edges that obey the triangle inequality [35]. While typically much smaller than the original network, the edges of the *metric backbone* are necessary and sufficient to compute all shortest paths in the original network. Our results show that it is indeed a primary subgraph for transmission dynamics. This is true for both the macro-dynamics of the overall times to infection, as previously suggested [44], and also for the micro-level dynamics captured by the time at which individuals get infected (see figure 2(b) and section 2.2). Specifically, even though the *metric backbone* induces a global delay in the spreading dynamics, it faithfully classifies the order in which individuals get exposed to the outbreak, and thus provides useful guidance for prioritizing control and mitigation strategies to disrupt epidemic spreading.

Interestingly, our results reveal that the impact on dynamics of the (semi-metric) edges that are not included in the *metric backbone* varies considerably. Indeed, effective reconstruction of outbreaks is shaped by a nontrivial interplay between semi-metric edge quantity and quality, by which we mean how much they break the triangle inequality, given by the semi-metric distortion parameter s^m (equation (6), Methods section). Edges with $s^m \gg 1$ are very far from contributing to any shortest path, whereas those with $s^m \approx 1$ link nodes with a distance almost the same as the shortest path (section 2.2)—in other words, the 'cost' of connecting two nodes directly via a low distortion edge is little more than via the (slightly shorter) indirect shortest path.

Our results demonstrate that removing high distortion edges hardly affects the dynamics regardless of the relative size of the backbone. In contrast, semi-metric edges with small distortion are important for recovering spreading dynamics. When there are more of these semi-metric edges, the size of the *metric backbone* becomes much more relevant for recovering the spreading dynamics accurately, whereby smaller backbones on their own do not recover dynamics as effectively as larger backbones (see figures 2(a), 3 and section 2.2). This suggests that availability of low distortion edges makes epidemic outbreaks robust by providing 'near shortest' transmission alternatives to shortest paths, whereas high distortion edges very rarely provide alternative paths in spreading dynamics and can be safely removed. Notice that it is possible to remove edges from the *metric backbone* itself and still preserve the distribution of shortest distances (or path lengths) of the original network [35]. This happens when there are alternative shortest paths with the exact same length between some node pairs. Importantly, the edges of all alternative shortest paths are kept in the *metric backbone* since they all obey the triangle inequality (because it is not a strict inequality). Our results show that those alternative equivalent paths should indeed be kept in sparsified networks as the *metric backbone* does, because they are not redundant for spreading dynamics as they have small semi-metric distortion.

These results indicate that measuring the semi-metric topology of complex networks, induced by distance functions that break the triangle inequality, is crucial for accurate reconstruction of spreading phenomena from sparsified networks. We thus developed and evaluated the SMDS to progressively sparsify weighted graphs by first removing the edges with the largest semi-metric distortion. Complementary to the unique sparsification given by identifying the metric backbone, SMDS can progressively reduce the entire network until its metric backbone, allowing for more or less aggressive sparsification. We compare SMDS with two sparsification methods by thresholding (neither of which has a principled limit): (1) using proximity weights, a local sparsification method that targets the weakest connections, and (2) using effective resistance, a global sparsification method that penalizes the existence of multiple paths connecting two nodes. We show that SMDS not only yields a more accurate reconstruction of macro and micro spreading dynamics, but also ensures network functionality by not breaking the network into disconnected components even while providing greater sparsification and preserving all shortest paths. More broadly, our study reveals that the distribution of semi-metric distortion values—especially the existence of many edges with small distortion—does play an important role in spreading dynamics. Because all semi-metric edges have zero edge betweenness, this result implies that new edge centrality measures that incorporate information about the semi-metric topology of weighted graphs would be useful to better capture edge importance.

Taken together, our findings demonstrate that the semi-metric topology of complex networks provides an algebraically-principled approach to network sparsification that is more effective than alternatives in recovering spreading dynamics. It should be noted that our results are restricted to widespread epidemic states. Under SI dynamics, the emergence of a global infected configuration is guaranteed and shortest paths appear as a natural driver for contagion processes. Likewise, our analysis of SIS and SIR dynamics is focused on outbreaks affecting a large fraction of the population. However, these models usually display localized epidemic outbreaks [55, 56], in which the importance of global communication over local connectivity for network sparsification remains an open question. Additionally, shortest paths have been reported to not capture some network diffusion processes, such as random walker dynamics, which can flow via longer but more likely network bypasses [57, 58]. Far from being limitations, these issues call for extending our study beyond shortest paths defined by distance functions constrained by the standard triangle inequality. We should consider general distance backbones induced by a generalized triangle inequality that can consider any measure of path length (not just the sum of distance weights) to characterize indirect transmission cost measures for other forms of spreading dynamics [35]. We are confident that both the methodology presented here and our results will motivate further study of the role of backbone subgraphs involved in driving and controlling network dynamics.

4. Methods

4.1. Metric backbone and semi-metric distortion

Let us assume that we have a system of N individuals whose contact associations are quantified by a proximity matrix $\mathbf{P}$, with entries p_{ij} denoting the strength of association between nodes (individuals) i and j—typically a normalized measure of how often the two individuals were together [44]. Obtaining the *metric backbone* involves the computation of shortest paths in the network; thus, we need a map between the proximity matrix $\mathbf{P}$ and a distance matrix $\mathbf{D}$, whose entries, d_{ij}, represent a notion of distance between nodes i and j, such that a small value denotes a strong association. Furthermore, this map should not affect the topology of $\mathbf{P}$, requiring it to be an isomorphism, such as:

$$d_{ij} = \frac{1}{p_{ij}} - 1\,, \tag{3}$$

with $d_{ij} \to \infty$ in case two individuals are never together [45]. This isomorphic relation is an important feature of the methodology, as it allows us to convert all key features of the semi-metric analysis to any weighted graph [35]. Thus, concepts such as shortest paths, the metric backbone, and semi-metric distortion introduced below are guaranteed to exist in weighted graphs whose edges denote proximity or strength, instead of distance or dissimilarity [44]. This easy conversion ensures that original meaning of edge weights is preserved, adding to the explainability power of the method.

Once the direct distances between nodes are defined, we can consider paths that connect nodes i and j indirectly: $\Gamma = i, k_1, \ldots, k_n, j$, where n is the number of intermediate nodes in the path. In the case of the metric backbone, the length of a path is computed as

$$d^{m}_{ij}(\Gamma) = d_{ik_1} + d_{k_1k_2} + \cdots + d_{k_nj}\,. \tag{4}$$

Note that we use the superscript m to indicate that all the measures in this manuscript are related to the metric algebraic space. Other choices for computing path length are possible, leading to generalized distance backbones [35], which are to be considered in future work. The shortest path metric length connecting both nodes (or shortest indirect distance) is in turn computed as $d^{T,m}_{ij} = \min(d_{ij}(\Gamma))$. Computing the shortest path metric length for all node pairs in $\mathbf{D}$ is known as the *all-Pairs Shortest Path problem* [59] and results in the metric closure matrix, denoted by $\mathbf{D}^{\mathrm{T,m}}$ [45]. Comparing the entries of this matrix with the corresponding finite entries of $\mathbf{D}$ reveals which edges obey or break the triangle inequality for any indirect path (via any number of indirect nodes): $d_{ij} \leqslant d^{T,m}_{ij}, \forall_{d_{ij}<\infty}$. In the first case, there are no shorter indirect paths, and the edge is referred to as *metric* and denoted and defined as $b_{ij} \equiv d^{T,m}_{ij} = d_{ij}$. However, in real-world networks there are typically many *semi-metric* edges that break this triangle inequality, observing $d_{ij} > d^{T,m}_{ij}$ instead [46]. Importantly, only the metric edges are necessary to compute all shortest paths (and sufficient to compute all shortest indirect distances), and thus define the *metric backbone* **B**, whose *relative size* is given by

$$\tau^m = \frac{|\{b_{ij}\}|}{|\{d_{ij} < \infty\}|}, \quad (5)$$

measuring the proportion of edges kept in this subgraph [35]. For unweighted graphs, $\tau^m = 1$, as all the direct edges represent the shortest path between their nodes. In contrast, for weighted graphs $\tau^m \leqslant 1$, depending on both the weights distribution and the specific structure of paths in each graph. Unless otherwise noticed, in this article when we refer to the triangle inequality, we mean the generalized case above which considers all possible indirect paths of any length (not simply directly connected triangles).

Since semi-metric edges do not participate in any shortest paths, they all have null edge betweenness (and associated centrality measure). However, they vary widely in their *semi-metric distortion* parameter given by:

$$s^m_{ij} = \frac{d_{ij}}{d^{T,m}_{ij}}, \quad (6)$$

which quantifies how much the edge between nodes *i* and *j* breaks the triangle inequality—in other words, how much shorter is the indirect distance (shortest path metric length) than the direct distance between them. For metric edges in the backbone, since the triangle inequality is satisfied, we naturally have $s^m_{ij} = 1$, while for semi-metric edges we have $s^m_{ij} > 1$.

Computing the semi-metric distortion values requires the computation of the all pairs shortest path problem which can be achieved via the Dijkstra algorithm with time complexity between $O(NE_m \log N)$ and $O(NE_m + N^2 \log N)$ depending on the implementation. Note that it depends on the number of edges in the metric backbone, not those in the overall network, since the later suffices to compute all shortest paths. Although not implemented, another possibility to compute the semi-metric distortion is by element-wise division of the entries in the distance adjacency matrix by the corresponding entries in the distance closure adjacency matrix. The latter, could provide a faster approximate implementation with latest algorithmic improvements in matrix operations [60], which is left as future work. On the other hand, if one is interested only in the metric backbone, then heuristic algorithms can achieve a time complexity as low as $O(NE)$ [61].

We refer the reader to [35] for a more exhaustive definition and study of the metric and other distance backbones, including the observed values of these parameters in networks across many domains.

4.2. Construction of synthetic networks

The first step to construct the synthetic networks used in this manuscript is to fix their metric backbone. To do so, we consider an undirected weighted network with N nodes and E_m edges, whose proximity values are captured in the matrix **B**. This initial subgraph has to satisfy all the constraints characterizing the *metric backbone* of a given network. Namely, it must have a single connected component and all its edges must be metric, i.e. they must represent the shortest path connecting their nodes. The next step involves mapping the proximity values to distances via equation (3) and computing the metric closure matrix $\mathbf{D}^{\mathrm{T,m}}$, thus obtaining the length of the shortest path connecting every pair of nodes in the network. Note that our method does not alter the structure of shortest paths in the network; therefore $\mathbf{D}^{\mathrm{T,m}}$ constitutes the metric closure matrix of the final synthetic network.

On top of the metric backbone, we add E_{sm} edges to tune its relative size τ^m compared to the total size of the constructed network. If $E = E_m + E_{sm}$ denotes the total number of edges in the constructed network, the relative size fulfills $E_m = \tau^m E$. Therefore, to fix a specific value τ^m, the number of semi-metric edges to be added is:

$$E_{sm} = \frac{1-\tau^m}{\tau^m} E_m. \quad (7)$$

These E_{sm} edges are chosen randomly within the set of edges not present in the *metric backbone* **B**. We choose them randomly as this is an agnostic way of constructing simple synthetic networks to unveil the limitations of the metric backbone. Note that $\tau^m \in \left[\frac{2E_m}{N(N-1)}, 1\right]$, where the lower bound corresponds to including all missing semi-metric edges to obtain a fully-connected network.

Once we fix the relative size of the backbone, we move to tuning the distortion of the semi-metric edges in the network. To do so, for each added edge, we sample its semi-metric distortion value s^m_{ij} from the target distribution $P(s^m)$ and assign the individual distance d_{ij} of the new link following equation (6). Therefore,

$$d_{ij} = s^m_{ij} d^{T,m}_{ij}, \quad (8)$$

which are eventually transformed into proximity values by using equation (3), yielding:

$$p_{ij} = \left(d_{ij} + 1\right)^{-1} \ . \tag{9}$$

4.3. SI dynamics

We focus on the SI compartmental model as a proxy for spreading dynamics due to its simplicity. In the SI model, there are only two states (compartments) available for each individual: Susceptible and Infected. A susceptible host i becomes infected when interacting with one infectious counterpart j at a rate βp_{ji}. Once nodes are infected, they remain in this state over the entire dynamics. Following those update rules, we perform agent-based simulations of the model. Therefore, the SI dynamics is entirely characterized by the distribution of times at which each agent i becomes infected, t_i. Note that, as shown in [44], β is not a relevant parameter to assess the *metric backbone* performance, as it just represents a global redefinition of the timescale of the spreading process. Throughout the manuscript, we fix $\beta = 0.5$ and perform all the simulations considering a single individual initially infected, constituting the seed of the infection. Generally, we characterize the outbreak by the time at which half of the population is infected, denoted by $t_{1/2}$. This value is obtained for 50 different initial seeds, to smooth out possible biases introduced by the origin of the outbreak in our analysis, and by averaging the results of 200 realizations for each seed.

One of the benefits of network sparsification methods which we highlight in this work is when they are able to capture those dynamical metrics (infection timing and order) in a reduced graph representation, since this would save time in those agent-based simulation of epidemic spreading models. For example, the machine runtime to measure $t_{1/2}$ using the *metric backbone* can reach down to 15% of the time it takes to measure $t_{1/2}$ in the original network (40 min instead of 4 h) as shown in table S2. In naturally sparser networks such as the c-elegans brain [62] and airports traffic [63], the runtime in the *metric backbone* is comparable to the original network but they bring clarity to network visualizations.

4.4. Effective resistance

One of the sparsification methods with which we compare the proposed SMDS is the removal of connections following the effective resistance edge ranking [41]. The effective resistance between two nodes i and j denoted by R^e_{ij} captures their global exchange of information through all the different paths connecting them in the network. Mathematically, one computes the effective resistance by:

$$R^e_{ij} = \left(e_i - e_j\right)^T L^\dagger \left(e_i - e_j\right) \ , \tag{10}$$

where $L^\dagger$ represents the Moore–Penrose inverse of the Laplacian matrix of the network and $\vec{e}$ the elements of the canonical basis. The effective resistance has proven to remove connections while preserving spreading SIR dynamics. Specifically, one can define the probability of keeping the edge connecting nodes i and j, p^R_{ij}, as [41]

$$p^R_{ij} \propto p_{ij} R^e_{ij} \ . \tag{11}$$

Thresholding the network according to the former probabilities prevents from isolating nodes, as $p^R_{ij} = 1$ when the edge (i,j) represents the single path connecting both nodes. As more paths becomes available, the former value becomes smaller, thus penalizing redundancy of information flow among nodes. To preserve SI dynamics, we must then remove those edges with lowest effective resistance values, as less relevant transmission pathways are hampered following their removal. Note that the ordering for edge removal according to their effective resistance values is computed only once considering the original network as a reference. The time complexity for measuring the effective resistance varies between $O(N^3)$ and $O(E \log N \log(1/\epsilon))$ depending on the implementation and approximation accuracy ϵ, as it is required to find the pseudo-inverse of the Laplacian matrix. Note that it depends on the number of edges in the overall network. It becomes comparable with the time complexity of SMDS for an accuracy of $\epsilon = \mathrm{e}^{-\tau N}$, where τ is the size of the backbone; otherwise, effective resistance can run about 4 times faster than the semi-metric distortion computation in finite size networks due to recent improvements in matrix operations [60].

4.5. Empirical networks

This study also uses 16 networks that were obtained from experimental data from social contacts (7), transportation (7), and nervous systems (2). Their properties, including the distribution of proximity values, the relative size of the metric backbone, and the semi-metric distortion distribution, are available in supplementary figure S1 and table S1.

All social contact networks considered were previously studied in [44], and are available in the SocioPatterns Database from their original studies [6, 47, 64–67]. Proximity edge strength values are obtained via proximity sensors that measure the time spent by individuals in the vicinity of each other. Specifically, we tally the number of 20 min intervals when each pair of individuals i and j were in direct contact with each other: r_{ij}. From this measurement, we compute the proximity matrix **P** defined in section 4.1 according to the Jaccard measure [68] for each entry:

$$p_{ij} = \frac{r_{ij}}{r_{ii} + r_{jj} - r_{ij}} , \tag{12}$$

where r_{ii} tallies the total number of 20 min intervals individual i was present in the experiment. The isomorphic distance matrix **D** is obtained via the equation (3).

Six transportation networks are similarly built from mobility data from core-based statistical areas (CSBA) in the state of New York in the United States (US). Proximity and distance matrices are built in the same way as the social contact networks via equations (3) and (12), but where r_{ij} tallies the number of trips between ZIP codes i and j, which are cast as nodes. A seventh transportation network is included to characterize air traffic between the 500 busiest commercial airports in the US during 2002 [63]. It is built in the same manner, with r_{ij} tallying the number of available airplane seats between airports i and j.

Two nervous system networks are also studied: a *human connectome*network mapping the strength of connections across 66 brain regions of interest [69], and a the neural network of the *Caenorhabditis elegans* worm (c-elegans) [62]. Proximity and distance matrices are built in the same way as the social contact networks via equations (3) and (12), but where r_{ij} tallies the number of streamlines, identified via diffusion spectrum imaging, per region volume between brain regions of interest i and j, using data from the human connectome extracted by [69], and the number of gap junctions between pairs of neurons i and j in the c-elegans nervous system network [62].

Data availability statement

The data that support the findings of this study are openly available at the following URL/DOI: https://github.com/CASCI-lab/SemiMetric_EpiSpread.

Acknowledgments

We thank Rion Brattig Correia and the members of the CASCI lab for useful discussions about this project. This work was funded by NIH National Library of Medicine Program Grant 01LM011945-01 and by the European Union's Horizon Europe research and innovation programme under Grant Agreement No. 101186558 for the 'Strategic Integration of Complex Networks and Systems for Advancing Biomedical Research' Project to LMR and by Portuguese national funds through the FCT - Foundation for Science and Technology, I.P., under the project 2022.09122.PTDC (https://doi.org/10.54499/2022.09122.PTDC) to LMR and FXC. DS-P was funded by MCIN/AEI/10.13039/501100011033 and the European Union 'NextGenerationEU'/PRTR' through Grants JDC2022-048339-I and PID2021-128005NB- C21. The funders had no role in any stage of the study design to the manuscript preparation, nor decision to publish.

ORCID iDs

David Soriano-Paños 0000-0002-6388-4056
Felipe Xavier Costa 0000-0001-6675-0355
Luis M Rocha 0000-0001-9402-887X

References

[1] Keeling M J and Eames K T D 2005 *J. R. Soc. Interface* **2** 295
[2] Morris M 2004 *Network Epidemiology: A Handbook for Survey Design and Data Collection* (Oxford University Press)
[3] Metcalf C J E, Morris D H and Park S W 2020 *Science* **369** 368
[4] Pagel C and Yates C A 2022 *BMJ* **378** e070615
[5] Blondel V D, Decuyper A and Krings G 2015 *EPJ Data Sci.* **4** 1
[6] Stehlé J *et al* 2011 *PLoS One* **6** e23176
[7] Gilbert M *et al* 2020 *Lancet* **395** 871
[8] Brockmann D and Helbing D 2013 *Science* **342** 1337
[9] Hazarie S, Soriano-Paños D, Arenas A, Gómez-Gardeñes J and Ghoshal G 2021 *Commun. Phys.* **4** 191
[10] Shea K *et al* 2023 *Proc. Natl Acad. Sci.* **120** e2207537120
[11] Meyers L A, Newman M, Martin M and Schrag S 2003 *Emerg. Infect. Dis.* **9** 204
[12] Rossa F D *et al* 2020 *Nat. Commun.* **11** 5106
[13] Rosenblatt S F, Smith J A, Gauthier G R, Hébert-Dufresne L and Tarnita C E 2020 *PLoS Comput. Biol.* **16** e1007897
[14] Mones E, Stopczynski A, Pentland A, Hupert N and Lehmann S 2018 *J. R. Soc. Interface* **15** 20170783

[15] Wei X, Zhao J, Liu S and Wang Y 2022 *Sci. Rep.* **12** 5550
[16] Reyna-Lara A, Soriano-Paños D, Arenas A and Gómez-Gardeñes J 2022 *Chaos Solitons Fractals* **158** 112012
[17] Ndeffo Mbah M L and Gilligan C A 2011 *PLoS One* **6** e24577
[18] Zhu X, Liu Y, Wang S, Wang R, Chen X and Wang W 2021 *Appl. Math. Comput.* **411** 126531
[19] Ciaperoni M, Galimberti E, Bonchi F, Cattuto C, Gullo F and Barrat A 2020 *Sci. Rep.* **10** 12529
[20] Wen S, Jiang J, Liu B, Xiang Y and Zhou W 2017 *J. Netw. Comput. Appl.* **78** 288
[21] Chung N N, Chew L Y, Zhou J and Lai C H 2012 *Europhys. Lett.* **98** 58004
[22] Liang G, Cui X and Zhu P 2023 *Front. Phys.* **11** 1164847
[23] Matamalas J T, Arenas A and Gómez S 2018 *Sci. Adv.* **4** eaau4212
[24] De Domenico M, Allegri L, Caldarelli G, d'Andrea V, Di Camillo B, Rocha L M, Rozum J, Sbarbati R and Zambelli F 2024 *npj Digit. Med.* **8** 37
[25] Rocha L M 2022 *Bioinformatics* **38** 3674
[26] Serrano M A, Boguná M and Vespignani A 2009 *Proc. Natl Acad. Sci.* **106** 6483
[27] Tumminello M, Aste T, Di Matteo T and Mantegna R N 2005 *Proc. Natl Acad. Sci.* **102** 10421
[28] Yan X, Jeub L G S, Flammini A, Radicchi F and Fortunato S 2018 *Phys. Rev. E* **98** 042304
[29] Radicchi F, Ramasco J J and Fortunato S 2011 *Phys. Rev. E* **83** 046101
[30] Marcaccioli R and Livan G 2019 *Nat. Commun.* **10** 745
[31] Gemmetto V, Cardillo A and Garlaschelli D 2017 (arXiv:1706.00230)
[32] Imre M, Tao J, Wang Y, Zhao Z, Feng Z and Wang C 2020 *Comput. Graph.* **87** 89
[33] Hermsdorff G B and Gunderson L 2019 *Adv. Neural Inf. Process Syst.* **32** (available at: http://papers.nips.cc/paper/8989-a-unifying-framework-for-spectrum-preserving-graph-sparsification-and-coarsening)
[34] Spielman D A and Srivastava N 2008 *Proc. 40th Annual ACM Symp. on Theory Comput.* pp 563–8
[35] Simas T, Correia R B and Rocha L M 2021 *J. Complex Netw.* **9** cnab021
[36] Costa F X, Correia R B and Rocha L M 2023 The distance backbone of directed networks *Complex Networks and Their Applications XI* vol 2, ed H Cherifi, R N Mantegna, L M Rocha, C Cherifi and S Micciche (*Studies in Computational Intelligence* vol 1078)(Springer) pp 135–47
[37] Wilson K G and Kogut J 1974 *Phys. Rep.* **12** 75
[38] García-Pérez G, Boguñá M and Serrano M A 2018 *Nat. Phys.* **14** 583
[39] Villegas P, Gili T, Caldarelli G and Gabrielli A 2023 *Nat. Phys.* **19** 1–6
[40] Swarup S, Ravi S, Mahmud M H and Lum K 2016 Identifying core network structure for epidemic simulations *Twelfth International Workshop on Agents and Data Mining Interaction (ADMI-2016)* (*Singapore*)
[41] Mercier A, Scarpino S and Moore C 2022 *PLOS Comput. Biol.* **18** e1010650
[42] Tolić D, Kleineberg K-K and Antulov-Fantulin N 2018 *Sci. Rep.* **8** 6562
[43] Gautreau A, Barrat A and Barthélemy M 2007 *J. Stat. Mech.* L09001
[44] Correia R B, Barrat A and Rocha L M 2023 *PLOS Comput. Biol.* **19** e1010854
[45] Simas T and Rocha L M 2015 *Netw. Sci.* **3** 227
[46] Rocha L M 2002 *Soft Computing Agents: A New Perspective for Dynamic Information Systems*, ed V Loia (IOS Press) pp 137–63 (arXiv:cs/0309013)
[47] Toth D J, Leecaster M, Pettey W B, Gundlapalli A V, Gao H, Rainey J J, Uzicanin A and Samore M H 2015 *J. R. Soc. Interface* **12** 20150279
[48] Pastor-Satorras R, Castellano C, Van Mieghem P and Vespignani A 2015 *Rev. Mod. Phys.* **87** 925
[49] Balcan D, Colizza V, Gonçalves B, Hu H, Ramasco J J and Vespignani A 2009 *Proc. Natl Acad. Sci.* **106** 21484
[50] Marques-Pita M and Rocha L M 2013 *PLoS One* **8** e55946
[51] Gates A J, Correia R B, Wang X and Rocha L M 2021 *Proc. Natl Acad. Sci.* **118** e2022598118
[52] Gates A J and Rocha L M 2016 *Sci. Rep.* **6** 24456
[53] Costa F X, Rozum J C, Marcus A M and Rocha L M 2023 *Entropy* **25** 374
[54] Manicka S, Marques-Pita M and Rocha L M 2022 *J. R. Soc. Interface* **19** 20210659
[55] Silva D H and Ferreira S C 2021 *J. Phys. Complex.* **2** 025011
[56] Cota W, Ferreira S C and Odor G 2016 *Phys. Rev. E* **93** 032322
[57] Estrada E, Gómez-Gardeñes J and Lacasa L 2023 *Proc. Natl Acad. Sci.* **120** e2305001120
[58] Lella E and Estrada E 2020 *Netw. Neurosci.* **4** 1007
[59] Zwick U 2002 *J. ACM (JACM)* **49** 289
[60] Koutis I, Levin A and Peng R 2012 *29th Int. Symp. on Theoretical Aspects of Computer Science (STACS 2012)* (Schloss Dagstuhl–Leibniz-Zentrum für Informatik) pp 266–77
[61] Kalavri V, Simas T and Logothetis D 2016 *Proc. VLDB Endowment* **9** 672
[62] Watts D J and Strogatz S H 1998 *Nature* **393** 440
[63] Colizza V, Pastor-Satorras R and Vespignani A 2007 *Nat. Phys.* **3** 276
[64] Salathé M, Kazandjieva M, Lee J W, Levis P, Feldman M W and Jones J H 2010 *Proc. Natl Acad. Sci.* **107** 22020
[65] Vanhems P, Barrat A, Cattuto C, Pinton J-F, Khanafer N, Régis C, Kim B-a, Comte B and Voirin N 2013 *PLoS One* **8** e73970
[66] Mastrandrea R, Fournet J, Barrat A and Viboud C 2015 *PLoS One* **10** e0136497
[67] Génois M and Barrat A 2018 *EPJ Data Sci.* **7** 11
[68] Jaccard P 1901 *Bull. Soc. Vaudoise Sci. Nat.* **37** 241
[69] Hagmann P, Cammoun L, Gigandet X, Meuli R, Honey C J, Wedeen V J and Sporns O 2008 *PLoS Biol.* **6** e159