Backbone Reconstruction in Temporal Networks from Epidemic Data

Francesco Vincenzo Surano[1,2], Christian Bongiorno[1,3], Lorenzo Zino[2], Maurizio Porfiri[2,*] and Alessandro Rizzo[1†]
[1]*Department of Electronics and Telecommunications, Politecnico di Torino, Turin, Italy*
[2]*Department of Mechanical and Aerospace Engineering,*
New York University Tandon School of Engineering, Brooklyn NY, USA
[3]*Laboratoire de Mathématiques et Informatique pour les Systèmes Complexes,*
CentraleSupélec, Université Paris Saclay, Gif-sur-Yvette, France
(Dated: July 12, 2019)

Many complex systems are characterized by time-varying patterns of interactions. These interactions comprise strong ties, driven by dyadic relationships, and weak ties, based on node-specific attributes. The interplay between strong and weak ties plays an important role on dynamical processes that could unfold on complex systems. However, seldom do we have access to precise information about the time-varying topology of interaction patterns. A particularly elusive question is to distinguish strong from weak ties, on the basis of the sole node dynamics. Building upon rigorous analytical results, we propose a statistically-principled algorithm to reconstruct the backbone of strong ties from epidemic data, consisting of the health state of individuals over time. Our method is numerically validated over a range of synthetic datasets, encapsulating salient features of real-world systems. Motivated by compelling evidence, we propose the integration of our algorithm in a targeted immunization strategy that prioritizes influential nodes in the inferred backbone. Though Monte Carlo simulations on synthetic networks and a real-world case study, we demonstrate the viability of our approach.

I. INTRODUCTION

In the last few decades, network science has experienced significant developments, providing researchers with an array of powerful tools to represent and analyze complex biological, social, and technological systems [1]. Besides improving our knowledge on the very structure of complex systems, network science has contributed new paradigms to study dynamical processes unfolding on a complex system. These paradigms have shed light on the intertwining between structure and dynamics in the spread of epidemic diseases [2], diffusion of innovation [3], and opinion formation [4].

Empirical studies suggest that patterns of interactions between nodes in many complex networks evolve ceaselessly in time [5, 6]. These interactions can be categorized into two main classes [7]. One class corresponds to interactions that are recurrently formed between node pairs, following dyadic relationships that are called *strong ties* [8]. Interactions in the workplace or family ties belong to this class, which forms the *backbone* of the network [9, 10]. The second class encompasses interactions that are based on features of the nodes, which are not attributable to dyadic ties with other nodes. For instance, interactions among people queuing in a line or sitting on a plane belong to this class, whereby interactions are triggered by individual attributes such as extroversion in talking to strangers. These relationships are called *weak ties* [8]. Strong and weak ties concur in shaping the dynamic behavior of complex networks [11–13].

Activity driven networks (ADNs) have emerged as a valuable framework for temporal networks [14], allowing for modeling the co-evolution of the network structure and the unfolding nodal dynamics at comparable time-scales. The temporal nature of the network is captured through a single parameter that measures the node propensity to generate interactions. The distribution of this parameter, called *activity*, can be inferred from real-world data [14]. The potential of ADNs has been demonstrated through the study of several network problems, including epidemics [15–19], diffusion of innovation [20], opinion formation [21], and percolation [22].

In their fundamental incarnation, ADNs are an ideal tool to model weak ties, whereby the whole process of network assembly is driven by a node-specific attribute, the activity. Routed ADNs (RADNs) have been recently proposed to include strong ties within the ADN paradigm [23, 24]. In this model, temporal connections are wired according to a stochastic rule that encapsulates both the topological information of strong ties and the unstructured connections of weak ties. RADNs share similarities with other approaches to include strong ties in ADNs, such as the superimposition of a static network [25, 26], and the inclusion of memory mechanisms in the link wiring process [27, 28].

The use of RADNs in epidemiological studies rely on accurate knowledge of the activity distribution and the topology of the backbone. While activities can be estimated following the literature on ADNs [14, 29], the inference of the backbone of strong ties remains an open challenge. Particularly elusive is the problem of distinguishing strong from weak ties using observations of the node dynamics, which is typically the only knowledge

[*] Also at Department of Biomedical Engineering, New York University Tandon School of Engineering, Brooklyn NY, USA; mporfiri@nyu.edu

[†] Also at Office of Innovation, New York University Tandon School of Engineering, Brooklyn NY, USA; alessandro.rizzo@polito.it

available in real epidemiological settings [30].

In the technical literature, the problem of link reconstruction and prediction has been studied from a variety of angles, mostly relying on the direct observations of contacts [31–33]. Dealing with observations of nodal dynamics, several methods have been proposed to reconstruct patterns of interactions [34], including the use of similarity [35], information theory [36], belief propagation [37], likelihood maximization [38], compressed sensing [39], optimization [40], and nonparametric Bayesian methods [41]. However, these methods are of limited use when strong and weak ties coexist, thereby presently challenging the inference of backbone networks from observations of node dynamics.

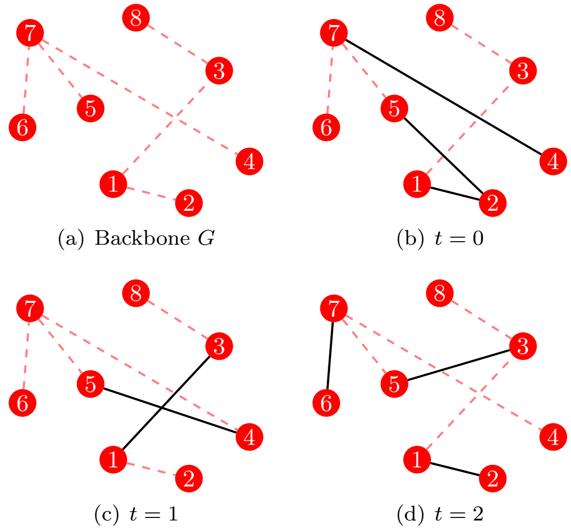
Drawing inspiration from [42, 43], here we design a backbone detection algorithm that identifies strong ties from node dynamics, in the form of empirical data about the spread of a disease. Our algorithm is based on the intuition that strong ties should leave a distinguishable footprint on the temporal evolution of an epidemic outbreak. We analytically characterize such a footprint in terms of the probability for a node to contract the disease, given knowledge about the health state of other nodes. Building upon this analytical result, we formulate a statistically-principled algorithm to reconstruct the backbone topology. An extensive performance analysis is carried out by means of numerical simulations to demonstrate the effectiveness of the algorithm and identify potential limitations. Finally, we demonstrate the possibility of implementing the algorithm to inform immunization strategies that target influential nodes of the backbone. The effectiveness of the proposed technique is evaluated through Monte Carlo simulations both on synthetic networks and real-world data of face-to-face interactions in a high school [44].

II. MATHEMATICAL BACKGROUND

We provide mathematical details of the models herein used to study temporal networks with a backbone structure of strong ties, along with dynamical process.

A. Routed ADNs

We consider a network of n nodes, each belonging to the node set $V = \{1, \dots, n\}$. Temporal undirected links are represented through time-varying adjacency matrix $A_t \in \{0,1\}^{n\times n}$, where $t \in \mathbb{Z}_+$ is the discrete time index. The adjacency matrix is assembled so that $(A_t)_{ij} = 1$, if and only if node i is connected with node j at time t. We denote by N_t^i the set of other nodes to which i is connected at time t.



(a) Backbone G

(b) $t = 0$

(c) $t = 1$

(d) $t = 2$

FIG. 1. Illustration of a backbone network along with three consecutive realizations of an RADN. Red dashed links are the strong ties in the backbone, and black solid links are temporal links generated from nodes' activity.

Both strong and weak ties contribute to the evolution of A_t. Strong ties are described by an undirected and time-invariant adjacency matrix $G \in \{0,1\}^{n\times n}$. We indicate with d_i the degree of node i in the backbone network. Degrees are gathered in the degree vector $d \in \mathbb{N}^n$. Empirical evidence from real-world observations suggests that real-world backbones are often sparse [1] and nodes have bounded degree [45]. Without loss of generality, we assume that the backbone network does not contain isolated node, that is, $d_i \geq 1$, for all $i \in V$ [46].

Following [24], each node $i \in V$ is characterized by an activity parameter $a_i \in [0,1]$. At each time, node i activates with probability a_i and generates an undirected link with another node. The selection of which node to connect to is probabilistically dictated by a row-stochastic [47] matrix $P \in \mathbb{R}^{n\times n}_{\geq 0}$ such that

$$P = (1-\gamma)\,\frac{1}{n-1}J + \gamma\,\mathrm{diag}(d)^{-1}G, \tag{1}$$

where $\gamma \in [0,1]$ is a constant parameter and J is the $n\times n$ matrix of all ones, except the diagonal entries, which are set to 0. The generic entry P_{ij} represents the probability that i connects with j. The first term on the right hand side of (1) accounts for the weak ties, while the second summand models strong ties in the backbone. The parameter $\gamma \in [0,1]$ weights the role of strong versus weak ties in the formation of temporal links. When $\gamma = 0$, the model reduces to a standard ADN [14] such that strong ties are uninfluential; when $\gamma = 1$, the probability of a connection mirrors the adjacency matrix of the backbone network. A realization of an RADN is shown in Fig. 1.

To generate a temporal network from $t = 0$, up to time T, we implement the following steps:

1. the temporal adjacency matrix is initialized as $(A_t)_{ij} = 0$, for all $i, j \in V$;

2. each node $i \in V$ activates with probability a_i, independent of the others;

3. for each node i that is active, a node j is selected with probability P_{ij}, and we set $(A_t)_{ij} = (A_t)_{ji} = 1$; and

4. the time index t is incremented by 1; if $t \geq T$, the algorithm is terminated, otherwise it is resumed to step 1.

B. Susceptible–infected–susceptible model

We focus on a susceptible–infected–susceptible (SIS) epidemic model [48]. In an SIS model, each node of the network is characterized by a binary health state. Specifically, at time t, node $i \in V$ is either susceptible to the disease ($X_t^i = 0$) or infected ($X_t^i = 1$). At each time two contrasting mechanisms govern the evolution of the epidemic process: propagation and recovery. Each susceptible node can contract the disease through interactions with infected nodes.

The propagation of the disease may occur with probability $\lambda \in [0, 1]$ along each link of the RADN independently of the others, such that

$$\mathbb{P}(X_{t+1}^i = 1 \,|\, X_t^i = 0) = 1 - (1-\lambda)^{\sum_{j \in N_t^i} X_t^j}. \tag{2}$$

Following the recovery mechanism, instead, each node i that is infected at time t, recovers at time $t+1$ with probability $\mu \in [0, 1]$, becoming again susceptible to the epidemics.

III. BACKBONE DETECTION ALGORITHM

We present here the main technical contribution of this work, which consists of an algorithm to detect the backbone of strong ties in a temporal network from epidemic data. Our method is based on the exact computation of the probability of a node to contract the disease given the health states of other nodes. Building on the knowledge about neighbors, we are able to pinpoint the effect of the presence of strong ties through a statistical test.

A. Conditional probabilities for RADNs

Given two nodes, i and j, from the initial time to T, we define the following quantity:

$$\begin{aligned}\mathcal{P}_{j\to i} \;:=\; \frac{1}{T}\sum_{t=0}^{T-1}\Big[&\mathbb{P}(X_{t+1}^i = 1 \,|\, X_t^i = 0, X_t^j = 1)\\ &-\mathbb{P}(X_{t+1}^i = 1 \,|\, X_t^i = 0)\Big].\end{aligned} \tag{3}$$

The quantity $\mathcal{P}_{j\to i}$ summarizes the extent by which the infection of node i over the time window $1, \dots, T$ is explained by disease propagation from node j [49]. Intuition suggests that such a quantity is larger when i and j are connected by a strong tie, such that the infection of nodes connected by the backbone network will increase the chance of contracting the infection. For the considered RADN and a SIS process, mathematical analysis of this quantity, detailed in the Appendix, confirms this intuition.

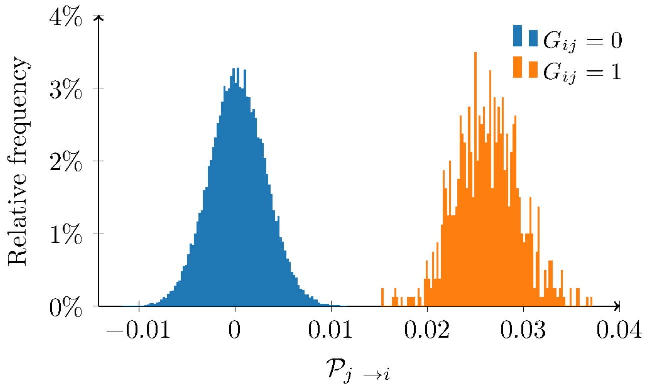


FIG. 2. Empirical estimation of $\mathcal{P}_{j\to i}$ in a realization of an RADN with $n = 200$ nodes, $\gamma = 0.95$, $\lambda = 0.9$, $\mu = 0.1$, and $a_i = 0.3$ for all nodes. The orange distribution relates to nodes that share a strong tie and the blue one to the opposite case. The backbone network is a 4-regular random graph. The network is simulated for $35,000$ time-steps. The figure suggests that conditioning on the state of node j affects the infection probability for nodes that share a strong tie with j, confirming our analytical results.

Specifically, we demonstrate that, in the asymptotic limit of large time-windows, if there is a strong tie between i and j, that is, $G_{ij} = 1$, then

$$\lim_{T\to\infty} \mathcal{P}_{j\to i} \geq \frac{\mu\gamma\lambda(1-\lambda)^{d_i - 1}}{e^3(1+\mu)}\left(\frac{a_i}{d_i} + \frac{a_j}{d_j} - \frac{\lambda a_i a_j}{d_i d_j}\right) > 0, \tag{4a}$$

almost surely, for any network size. On the other hand, if the nodes are disconnected in the backbone, that is, if $G_{ij} = 0$, we find that in the asymptotic limit of large networks,

$$\lim_{n\to\infty} \mathcal{P}_{j\to i} = 0. \tag{4b}$$

As a consequence, if the size of the network is sufficiently large, the probability that a node becomes infected is not influenced by the health state of another, unless they share a strong tie. Based on this analytical result, we construct our identification algorithm, which starts from empirical observations of the disease dynamics to detect strong ties.

Figure 2 compares the empirical estimation of $\mathcal{P}_{j\to i}$ for pairs of nodes that share (orange) or not (blue) a strong tie. These simulations validate our analytical results and suggest that $\mathcal{P}_{j\to i}$ is close to its asymptotic expressions in (4), also for a reasonable small population size and a limited observation window. In fact, while the empirical distribution of the entries of $\mathcal{P}_{j\to i}$ that correspond to strong ties (in orange) is shifted and bounded away from 0, the empirical distribution of the entries that do not correspond to strong ties is centered at 0. We notice that the two empirical distributions are well separated.

B. Statistical test

Building on our analytical results, we put forward a statistical method to determine the presence of a strong tie between the two nodes for a large network. To perform such a statistical analysis, for any pair of nodes i and j, we measure the following four quantities:

- the number of time-steps in which node i is susceptible, denoted as s_i;
- the number of transitions of node i from susceptible to infected, denoted as i_i;
- the number of time-steps in which node i is susceptible and node j is infected, denoted as n_{ij}; and
- the number of transitions of node i from susceptible to infected with node j being infected at the previous time, denoted as q_{ij}.

From the first two quantities, we compute the ratio $r_i = i_i/s_i$, which measures the sampling probability that a susceptible node i at time t becomes infected at $t+1$.

According to (4b), if i and j do not share a strong tie, then the probability that i contracts the infection should not be influenced by j, that is, q_{ij} should be a realization of a Bernoulli trial with expected value equal to $r_i n_{ij}$. We set this as null hypothesis of our statistical test, which is rejected if q_{ij} is significantly larger than $r_i n_{ij}$. We associate with the node pair a p-value, coming from the binomial cumulative distribution, equal to

$$\pi_{ij} = 1 - \sum_{h=0}^{q_{ij}-1} \binom{n_{ij}}{h} r_i^h \, (1-r_i)^{n_{ij}-h} \,. \tag{5}$$

This procedure generates a set of $n-1$ statistical tests for each node, that is, $n(n-1)$ tests, overall. Hence, a multiple comparison correction should be implemented to assess whether each one of the null hypotheses can be rejected. We adopt the Benjamini–Hochberg procedure to control the false discovery rate, which offers a less conservative criterion with respect to the standard Bonferroni criterion [50]. This method is implemented as follows.

First, we set the level of significance $\alpha \in [0,1]$. The quantity α measures the largest admissible probability that at least one of the null hypotheses is erroneously rejected and it is typically a small quantity, to ensure the test's significance. Then, the $n(n-1)$ p-values are sorted in ascending order as $\pi^{(1)} < \pi^{(2)} < \cdots < \pi^{((n-1)n)}$. Let L be the largest integer for which it holds $\pi^{(L)} < L\alpha/(n-1)n$. Then, the null hypothesis is rejected for all the pairs of nodes associated with a p-value smaller than $\pi^{(L)}$. If the null hypothesis is rejected for i and j, then we estimate that there is a link in the backbone network such that $\hat{G}_{ij} = \hat{G}_{ji} = 1$. We note that this is the step that requires the most computational effort, since the $n(n-1)$ p-values should be computed and sorted in ascending order. The algorithm can be implemented according to the pseudocode below.

Algorithm 1: Backbone detection algorithm

```
Data: empirical observations $r_i$, $n_{ij}$, $q_{ij}$, $\forall\, i,j \in V$
Result: estimation of the adjacency matrix $\hat{G}$
$\hat{G} \longleftarrow 0$;
for $i \in V,\ j \in V,\ j \neq i$ do
    compute $\pi_{ij}$ using (5);
sort $\pi_{ij}$ in ascending order $\pi^{(1)} \leq \pi^{(2)} \leq \dots$;
$L \longleftarrow \max\{k \in \mathbb{N} : \pi^{(L)} < L\alpha/(n-1)n\}$;
for $i \in V,\ j \in V,\ j \neq i$ do
    if $\pi_{ij} \leq \pi^{(L)}$ then
        $\hat{G}_{ij} \longleftarrow 1$;
        $\hat{G}_{ji} \longleftarrow 1$;
```

Examining more in depth the analytical results in (4a), we foresee some issues that might hinder the applicability of our algorithm, yielding a small value of $\mathcal{P}_{j\to i}$, even though a strong tie connecting i to j exists. In particular, this can occur in two cases. First, if both degrees d_i and d_j are large, such that the two nodes have a large degree centrality in the backbone network. Second, if both activities a_i and a_j are small. In the following, we present detailed numerical simulations with different parameter choices to demonstrate the accuracy of the algorithm.

IV. NUMERICAL VALIDATION

We validate our backbone detection algorithm on several synthetic datasets, to illustrate its applicability in real-world scenarios and identify potential limitations. These synthetic datasets consist of benchmark networks with $n = 200$ nodes, generated according to the RADN paradigm described in Section II A. We set $\gamma = 0.95$, when weighting temporal versus backbone contacts in (1). We consider different distributions for the activities and degree distribution of the backbone that follows a configuration model [1]. The epidemic process is simulated using the SIS model illustrated in Section II B with $\lambda = 0.9$ and $\mu = 0.1$. Unless otherwise specified, we set the significance level of the statistical test to $\alpha = 0.05$.

A. Homogeneous activity distribution and homogeneous backbone

We first examine the possibility of identifying regular networks of strong ties against weak ties generated us-

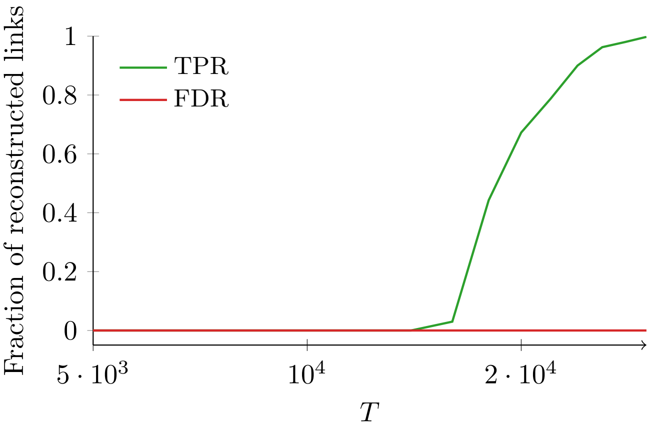


FIG. 3. Fraction of strong ties identified by our algorithm. The backbone is a 4-regular network with 200 nodes. The other parameters are $\gamma = 0.95$, $\lambda = 0.9$, $\mu = 0.1$, and $a_i = 0.3$, for all the nodes.

ing a common activity value for the all nodes. In this scenario, the backbone is chosen to be a 4-regular random network and the activity is equal to $a_i = 0.3$, for all $i \in V$.

In Fig. 3, we report the true positive rate (TPR), which is the fraction of links that the algorithm is able to correctly predict (green); and the false discovery rate (FDR), which is the ratio between the number of times it fails to properly identify a link and the number of links in the backbone (red). Perfect reconstruction is attained when the number of true positives is equal to the total number of positives (TPR= 1) and the number of false positives is close to zero (FDR= 0). The computations are carried out for different values of T, such that larger values of T imply access to a longer time window for the estimation of the probabilities of transitions in the algorithm.

For sufficiently large values of T, our algorithm is successful in exactly reconstructing the topology of the backbone. Choosing small values of T hampers the identification of links, but it rarely results into the identification of false positives (four false positives are overall identified in Fig. 3), such that we progressively improve the detection of strong ties, attributing a very small quantity of wrong links to the backbone. This is an important feature of the algorithm, whereby all the links it discovers can be relied upon with an extremely high confidence.

B. Heterogeneous activity distribution and homogeneous backbone

To better proxy a real-world setting, we release the assumption that all the nodes have the same activity. As a stepping stone, we consider the case in which nodes are randomly divided into two activity classes with 100 nodes each: low-activity nodes ($a_i = 0.2$) and high-activity nodes ($a_i = 0.8$). Similar to the previous analysis, the backbone is a 4-regular random network. To help teasing out the role of heterogeneity, we also simulate the scenarios in which all the nodes are either in the low- or high-activity classes.

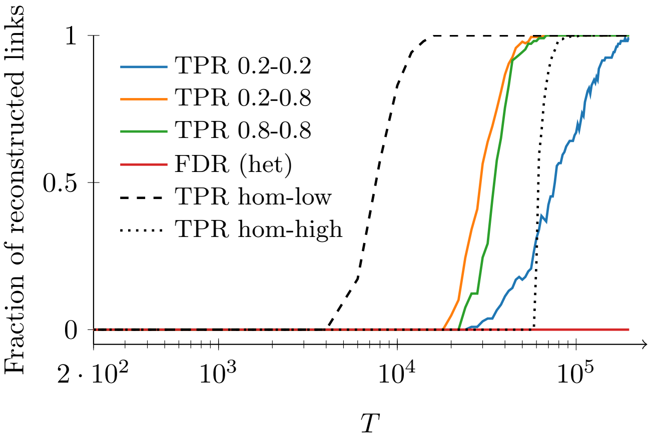


FIG. 4. Fraction of strong ties correctly identified by our algorithm for both heterogeneous and homogeneous activity distributions, and a regular network.. The backbone is a 4-regular network with $n = 200$ nodes. The other parameters are $\gamma = 0.95$, $\lambda = 0.9$, and $\mu = 0.1$. Three possibilities for the activity distribution are examined: all the nodes have the same activity $a_i = 0.2$ (hom-low, dashed), $a_i = 0.8$ (hom-high, dotted), and half the nodes have $a_i = 0.2$ and half have $a_i = 0.8$ (het, colored). For the last case of heterogeneous activities, TPR is reported with respect to links between nodes with low activity (blue), links between nodes of different activity (orange), and links between nodes with high activity (green). Only one FDR is reported for all the possibility, since they are indistinguishable (het, red).

Again, we examine the effect of T on true and false positives, with respect to the number of positives. Results in Fig. 4 confirm those from Fig. 3, whereby the fraction of correctly identified links increases with T and the fraction of misclassified links is always negligible. Comparing the three scenarios, we observe that large values of the activity have a negative effect on the performance of the algorithm. In fact, an increased observation window is required to detect strong ties in the homogeneous case with high activity, with respect to the scenario with low activity.

Heterogeneity further reduces performance, hampering the detection of strong ties between low-activity nodes. Even though networks with a heterogeneous activity distribution require a longer window to correctly detect all the strong ties, we observe that, for sufficiently large T, our algorithm is able to correctly reconstruct the backbone, with a negligible fraction of erroneous identifications. Overall, these results are in agreement with theoretical analysis in the Appendix, whereby decreasing the activities causes a reduction in the probability difference in (4a).

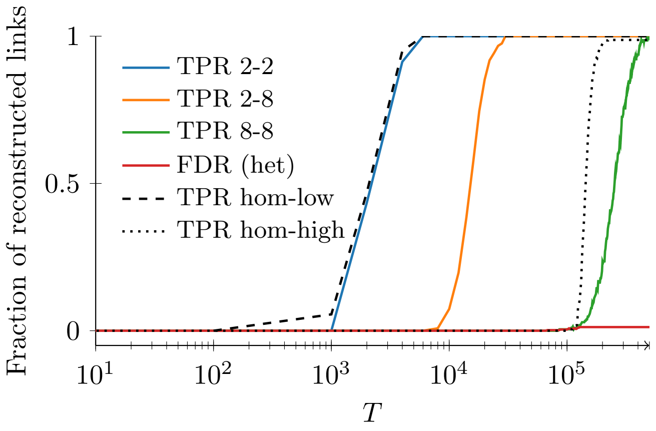


FIG. 5. Fraction of strong ties correctly identified by our algorithm for both heterogeneous and homogeneous backbones, and activity $a_i = 0.3$, for all the nodes. The other parameters are $n = 200$, $\gamma = 0.95$, $\lambda = 0.9$, and $\mu = 0.1$. Three possibilities for the backbone are examined: all the nodes have the same degree $d_i = 2$ (hom-low, dashed), $d_i = 0.8$ (hom-high, dotted), and half the nodes have $d_i = 2$ and half have $d_i = 8$ (het, colored). For the last case of heterogeneous degrees, TPR is reported with respect to links between nodes with low degree (blue), links between nodes of different degree (orange), and links between nodes with high degree (green). Only one FDR is reported for all the possibility, since they are indistinguishable (het, red).

C. Homogeneous activity distribution and heterogeneous backbone

Next, we examine a backbone where the degree of the nodes is not held constant throughout the network. Specifically, we consider a network in which nodes are partitioned into two classes of 100 nodes each with low- ($d_i = 2$) or high-degree ($d_i = 8$). To avoid confounding, we maintain the activity at a common value of $a_i = 0.3$, similar to results in Fig. 3. Once again, to facilitate the assessment of the effect of a heterogeneous degree distribution on the algorithm performance, we analyze two control cases in which the all the nodes have the same low- or high-degree.

Figure 5 illustrates the fraction of links predicted as a function of T for three considered settings. Consistent with our previous results, we observe that increasing the length of the observation steadily benefits the algorithm precision in inferring strong ties, as shown in Fig. 5. The number of false positives is always negligible, even for small values of T, confirming that the algorithm can be reliably utilized for backbone inference.

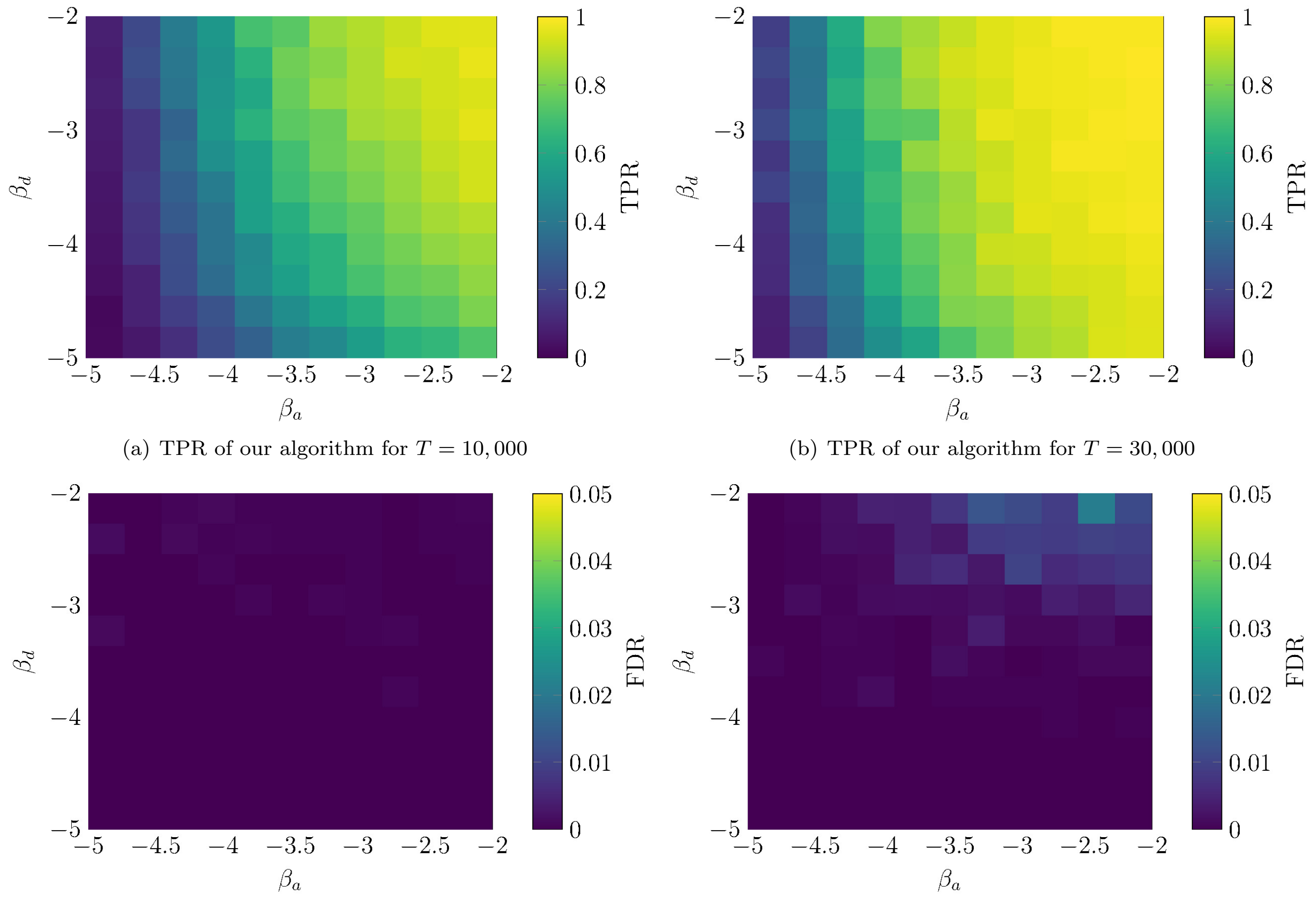
Comparing the two homogeneous cases of low- and high-degree distributions, we register an expected decrease in performance when dealing with higher degrees. In this case, the value of added knowledge regarding the state of health of one node is diluted by the presence of many other neighbors that could have triggered the infection. Analytical results in the Appendix provide a theoretical basis for this explanation, whereby increasing the values of the degree causes a reduction in the probability difference in (4a).

As one might expect, the performance of the algorithm toward the inference of the heterogeneous network is in between the two cases of homogeneous networks. To gain further insight into the relationship between topological features and successful reconstruction, we can isolate the specific links that are first detected by the algorithm for small values of T. In agreement with our analytical result in (4a), the links that require shorter observations are incident to low-degree nodes. These links encompass both strong ties between low-degree nodes and strong ties between nodes with high and low degrees that might exemplify dissortative structures of real networks [51, 52]. Longer time windows are required for detecting links that connect pairs of high-degree nodes.

D. Highly-heterogeneous activity distribution and backbone

To offer insight on the performance of our algorithm over a wider class of RADNs, we systematically examine a two-dimensional grid of salient parameters. We assume that both the activity and the degree distributions follow a power-law with exponents β_a and β_d, respectively. We vary each parameter from -5 to -2, which are representative of real-world scenarios [53]. Parameters are varied in 11 steps with cutoffs at 0.1 and 1 for the activity, and at 1 and $n - 1$ for the degree. We observe that smaller values of the exponent of a power-law yield distributions with a larger dispersion, in which most of the nodes have small activity (degree) and few have an extremely high activity (degree). Two different realizations are examined, one with $T = 10,000$ and $T = 30,000$, respectively. The weight γ is reduced to 0.5 to guarantee the spread of the epidemic diseases for all the choices of parameters investigated and the network size is increased to $n = 300$ to ensure the presence of high-degree (activity) nodes in the power-law distributions. The epidemic parameters are set as $\lambda = 0.9$ and $\mu = 0.1$, similar to the simulations in Section IV.

From Fig. 6, we recognize a marked effect of the parameters on the performance of our algorithm. For lower values of both parameters, β_a and β_d, our algorithm fails to identify the backbone, under-predicting the number of strong ties. This is in agreement with Figs. 4 and 5, which indicate that longer observation windows are re-



(a) TPR of our algorithm for $T = 10,000$

(b) TPR of our algorithm for $T = 30,000$

(c) FDR of our algorithm for $T = 10,000$

(d) FDR of our algorithm for $T = 30,000$

FIG. 6. TPR (a,b) and FDR (c,d) of our algorithm implemented on a network of $n = 300$ nodes for an observation window of $T = 10,000$ time-steps (a,c) or $T = 30,000$ time-steps (a,c). Both activities and degrees in the backbone follow power-law distributions with exponents β_a and β_d, respectively. Other parameters are set to $\lambda = 0.9$, $\mu = 0.1$, and $\gamma = 0.5$. Each point is an average of ten independent simulations.

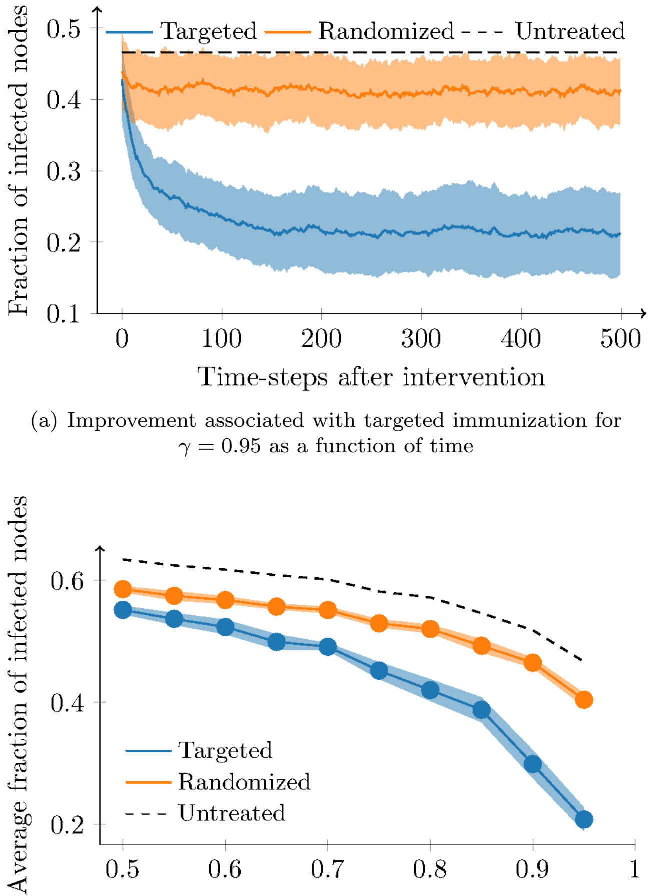
quired to infer the backbone when the RADN is dominated by high-degree and high-activity nodes. The best performance is attained for higher values of the two parameters. In this case, the algorithm correctly detects all the strong ties, with a very small quantity of false positives.

Comparing the results for $T = 10,000$ and $T = 30,000$, interestingly, β_a seems to have a stronger effect on performance than β_d, whereby at $T = 30,000$, the algorithm is able to detect most of the strong ties for small values of β_d but its performance is strained when examining small values of β_a. This confirms our preliminary observation from Fig. 4 that heterogeneity in the activity distribution hampers the detection of strong ties.

V. APPLICATION TO TARGETED IMMUNIZATION

In epidemiology, knowledge about the backbone network might offer valuable information about how diseases spread and which is the role played by individuals. In this vein, we conclude this paper by presenting an application of our algorithm to design a targeted immunization protocol. Our control strategy observes the disease spreading for a finite time-window to identify the backbone network, and then utilizes such an inference to prioritize immunization of nodes in the network according to a centrality criterion. Specifically, we immunize nodes according to decreasing values of their PageRank centrality [54]. By means of Monte Carlo numerical simulations, we evaluate the performance of the approach against a randomized immunization, where no information regarding the backbone is utilized.

Similar to the analysis in Section IV D, we examine a benchmark network with $n = 300$ nodes. The backbone is generated using a configuration model with power-law degree distribution of power $\beta_d = -3$ and cutoffs at 1 and $n - 1$. Activities are also drawn from a power-law distribution with exponent $\beta_a = -3$ and lower cutoff at 0.1. We consider an SIS epidemic with $\lambda = 0.9$ and



(a) Improvement associated with targeted immunization for $\gamma = 0.95$ as a function of time

(b) Improvement associated with targeted immunization for different values of γ

FIG. 7. Monte Carlo estimation over 100 runs of the effect of randomized (orange) and targeted (blue) immunization on the fraction of infected nodes. Dotted lines indicate the fraction of infected nodes in the absence of any immunization technique. In (a), we show the entire realizations for $\gamma = 0.95$. The solid line is the average, while the light band is one standard deviation. In (b), we compare the average fraction of infected nodes for different values of γ. Bands identify 95% confidence intervals. The lower panel illustrates the entire realizations for $\gamma = 0.95$. Other parameters are $n = 300$, $\beta_d = \beta_a = -3$, $\lambda = 0.9$, and $\mu = 0.1$.

$\mu = 0.1$. We run the model over a window of $50,000$ time-steps implementing our algorithm to identify the backbone. At this time, we execute two control strategies (targeted and randomized), with a number of interventions limited to 5% of the total number of nodes. We perform Monte Carlo simulations by averaging over 100 independent runs of the two control strategies.

The results of these simulations are summarized in Fig. 7. In Fig. 7(a), we compare the performance of the two immunization strategies for $\gamma = 0.95$, as in the numerical analysis in Section IV. While randomized immunization decreases the portion of infected nodes by 13%, targeted intervention decreases it by 55%, on average. The difference between these two strategies is statistically significant (p-value $\ll 0.0001$, according to a two-sample z-test) comparing the average fraction of infected individuals after the implementation of the immunization strategy, for 100 independent runs. In Fig. 7(b), instead, the comparison between the two techniques is conducted for different values of the parameter γ, spanning from 0.5 to 0.95 in steps of 0.05. Therein, we report the average fraction of infected nodes in the 500 time-steps that follow the application of the control strategy. Predictably, the larger the parameter γ, the stronger the improvement of the targeted immunization with respect to the randomized one. In fact, for small values of γ, the backbone has a marginal role on the link formation process, reducing the effect of targeted immunization exploiting the centrality measures in the backbone. However, the difference between the two strategies is statistically significant in all the performed simulations.

Encouraged by these promising results, we apply our targeted immunization technique to real-world face-to-face interactions measured through proximity sensors in a high school [44], available at [55]. The dataset comprises $188,508$ temporal links, generated over $T = 7,375$ time-steps among $n = 327$ nodes. We run an SIS epidemic model for half of the available dataset, starting from a fraction of one third of infected nodes, selected uniformly at random. Then, 5% of the nodes is immunized following either the randomized or the targeted strategy. By performing an extensive Monte Carlo simulation with $1,000$ runs, we compare the two strategies for different values of the epidemic parameters λ and μ. Figure 8 demonstrates that our immunization technique should always be preferred to randomized immunization, whereby, for most parameter choices, it outperforms randomized immunization.

VI. CONCLUSIONS

In this work, we have proposed an algorithm to unveil the backbone of strong ties in a temporal network from epidemic data. Building on analytical insight regarding the role of strong ties on the epidemic, we have put forward a statistically-principled approach to discover strong ties from empirical data. Extensive simulations have been performed to assess the effectiveness of the proposed technique, which has proved to be reliable in a variety of scenarios. Finally, we have examined the integration of the proposed algorithm in the solution of an important challenge in epidemiology, namely, targeted immunization during an outbreak. The main contributions of this work are: *i*) the analytical computation of the effect of strong ties on the infection probability for a susceptible–infected–susceptible epidemic model on routed activity driven networks; *ii*) the design of a backbone detection algorithm and its numerical validation; and *iii*) the implementation of a targeted immunization technique.

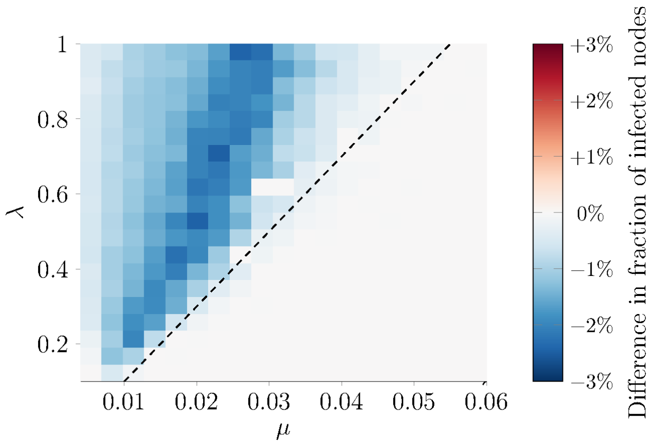


FIG. 8. Difference in the fraction of infected nodes after the immunization phase, between the randomized and the targeted strategy (color coded) in the high-school case study [55]. The dashed line represents the epidemic threshold [48], below which none of the nodes is infected at the onset of the immunization strategy. Darker blue areas identify parameter regions where targeted immunization has superior outcome. Each point is an average of 1,000 independent simulations.

The promising preliminary of our numerical analysis pave the way for several avenues of future research. We aim to rigorously assess the performance of our algorithm, as a function of the network size and the duration of the window of observation. In most real-world scenarios, it is not tenable to have access to the entire node set, thereby calling for methods to discover missing nodes, beyond links. Finally, our study on targeted immunization has demonstrated how information about the backbone can be leveraged to design effective control techniques that could steer the behavior of dynamical systems. Extending the framework to other disease models and mathematically proving performance bounds is the objective of future research.

ACKNOWLEDGMENTS

This work was partially supported by the National Science Foundation under grant number CMMI-1561134; the Compagnia di San Paolo, Torino, Italy; and the Italian Ministry of Foreign Affairs and International Cooperation under grant “Mac2Mic.” The authors acknowledge the SocioPatterns group.

Appendix A: Computation of the conditional probabilities

We compute the infection probability for node i at time instant t, for either the case in which we include or exclude knowledge about node j. Let $x_1,\dots,x_n$ be the state of the system at time t, then the RADN model indicates that

$$\mathbb{P}(X^i_{t+1}=1\,|\,X^i_t=0)=1-\prod_{k\in V\setminus\{i\}}(1-\lambda a_i P_{ik}x_k)\,(1-\lambda a_k P_{ki}x_k)\,. \tag{A1}$$

Upon conditioning on $X^i_t=1$, we factor the term associated with j out of the multiplication to obtain

$$\mathbb{P}(X^i_{t+1}=1\,|\,X^i_t=0,X^j_t=1)=1-(1-\lambda a_i P_{ij})\,(1-\lambda a_j P_{ji})\prod_{k\in V\setminus\{i,j\}}(1-\lambda a_i P_{ik}x_k)\,(1-\lambda a_k P_{ki}x_k)\,. \tag{A2}$$

First, we consider the case in which nodes i and j do not share a strong tie, that is $G_{ij}=G_{ji}=0$. In this case, from (1) we derive $P_{ij}=P_{ji}=(1-\gamma)/(n-1)$. We substitute P_{ij} and P_{ji} in (A1) and (A2), and we compute the limit for $n\to\infty$ of their difference as

$$\begin{aligned}&\lim\nolimits_{n\to\infty}\mathbb{P}(X^i_{t+1}=1\,|\,X^i_t=0,X^j_t=1)-\mathbb{P}(X^i_{t+1}=1\,|\,X^i_t=0)=\\&=\lim_{n\to\infty}\left[\left(1-\frac{\lambda(1-\gamma)a_i x_j}{n-1}\right)\left(1-\frac{\lambda(1-\gamma)a_j x_j}{n-1}\right)-\left(1-\frac{\lambda(1-\gamma)a_i}{n-1}\right)\left(1-\frac{\lambda(1-\gamma)a_j}{n-1}\right)\right]\times\\&\times\prod_{k\in V\setminus\{i,j\}}(1-\lambda a_i P_{ik}x_k)\,(1-\lambda a_k P_{ki}x_k)=0.\end{aligned} \tag{A3}$$

We note that (A3) is the generic summand of $\mathcal{P}_{j\to i}$ in (3), from which the claim in (4b) follows.

We now consider the case in which nodes i and j share a strong tie, that is, $G_{ij}=G_{ji}=1$. Similar to the previous analysis, from (1) we derive $P_{ij}=(1-\gamma)/(n-1)+\gamma/d_i$ and $P_{ji}=(1-\gamma)/(n-1)+\gamma/d_j$. Defining the neighborhood of node i in the backbone $N^i_G:=\{j\in V:G_{ij}=1\}$, we proceed specializing to the present case the difference between (A1) and (A2) at time t. Considering that $(1-1/x)^{x-1}\geq 1/e$, for any $x\geq 1$, and that $d_i\leq n-1$, for any

$i \in V$, we compute

$$\begin{aligned}
&\mathbb{P}(X^i_{t+1}=1\,|\,X^i_t=0, X^j_t=1)-\mathbb{P}(X^i_{t+1}=1\,|\,X^i_t=0)=\\
&=\left[\left(1-\lambda a_i x_j\left(\frac{\gamma}{d_i}+\frac{1-\gamma}{n-1}\right)\right)\left(1-\lambda a_j x_j\left(\frac{\gamma}{d_j}+\frac{1-\gamma}{n-1}\right)\right)-\left(1-\lambda a_i\left(\frac{\gamma}{d_i}+\frac{1-\gamma}{n-1}\right)\right)\left(1-\lambda a_j\left(\frac{\gamma}{d_j}+\frac{1-\gamma}{n-1}\right)\right)\right]\times\\
&\qquad\times\prod_{k\in V\setminus\{i,j\}}(1-\lambda a_i P_{ik}x_k)\,(1-\lambda a_k P_{ki}x_k)\\
&\geq \lambda\gamma(1-x_j)\left(\frac{a_i}{d_i}+\frac{a_j}{d_j}-\lambda\gamma\frac{a_i a_j}{d_i d_j}\right)\prod_{k\in N^i_G\setminus\{j\}}(1-\lambda a_i P_{ik}x_k)\,(1-\lambda a_k P_{ki}x_k)\prod_{h\notin N^i_G\cup\{i\}}(1-\lambda a_i P_{ih}x_h)\,(1-\lambda a_h P_{hi}x_h)\\
&\geq \lambda\gamma(1-x_j)\left(\frac{a_i}{d_i}+\frac{a_j}{d_j}-\lambda\gamma\frac{a_i a_j}{d_i d_j}\right)\prod_{k\in N^i_G\setminus\{j\}}\left(1-\frac{1}{d_i}\right)(1-\lambda)\prod_{h\notin N^i_G\cup\{i\}}\left(1-\frac{1}{n-1}\right)\left(1-\frac{1}{n-1}\right)\\
&\geq \frac{\lambda\gamma(1-\lambda)^{d_i-1}}{e^3}(1-x_j)\left(\frac{a_i}{d_i}+\frac{a_j}{d_j}-\lambda\gamma\frac{a_i a_j}{d_i d_j}\right):=F(x_j),
\end{aligned}\tag{A4}$$

where the bounding function $F(x_j)$ is such that $F(1)=0$, and $F(0)>0$, for any $\gamma>0$.

We now focus on the variable X^j_t. According to the SIS dynamics described in II B, X^j_t changes from 1 to 0 with probability equal to μ, while the probability of switching from 0 to 1 depends on the health state of the other nodes, but is obviously bounded from above by 1. Hence, the frequency of $X^j_t=0$ converges almost surely to at least $\mu/(1+\mu)$ for $T\to\infty$. Hence, using (A4) and the definition of $\mathcal{P}_{j\to i}$ in (3), the latter quantity can be bounded from below as follows:

$$\begin{aligned}
\lim_{T\to\infty}\mathcal{P}_{j\to i} &= \lim_{T\to\infty}\frac{1}{T}\sum_{t=0}^{T-1}\left[\mathbb{P}(X^i_{t+1}=1\,|\,X^i_t=0,X^j_t=1)-\mathbb{P}(X^i_{t+1}=1\,|\,X^i_t=0)\right]\geq\lim_{T\to\infty}\frac{1}{T}\sum_{t=0}^{T-1}F(X^i_t)\\
&=\lim_{T\to\infty}\frac{1}{T}\sum_{t\in\{0,\dots,T-1\}:X^i_t=0}F(0)\geq\frac{\mu}{1+\mu}F(0)\geq\frac{\mu\lambda\gamma(1-\lambda)^{d_i-1}}{e^3(1+\mu)}\left(\frac{a_i}{d_i}+\frac{a_j}{d_j}-\frac{\lambda a_i a_j}{d_i d_j}\right)>0.
\end{aligned}\tag{A5}$$

[1] M. E. Newman, SIAM Rev. **45**, 167 (2003).

[2] R. Pastor-Satorras, C. Castellano, P. Van Mieghem, and A. Vespignani, Rev. Mod. Phys. **87**, 925 (2015).

[3] A. Montanari and A. Saberi, Proc. Nat. Acad. Sci. USA **107**, 20196 (2010).

[4] P.-P. Li, D.-F. Zheng, and P. M. Hui, Phys. Rev. E **73**, 056128 (2006).

[5] P. Holme and J. Saramäki, Phys. Rep. **519**, 97 (2012).

[6] P. Holme, Eur. Phys. J. B **88** (2015).

[7] M. S. Granovetter, Am. J. Sociol. **78**, 1360 (1973).

[8] N. E. Friedkin, Soc. Netw. **3**, 273 (1982).

[9] V. Gemmetto, A. Cardillo, and D. Garlaschelli, arXiv preprint: 1706.00230 (2017).

[10] J.-P. Onnela, J. Saramäki, J. Hyvönen, G. Szabó, D. Lazer, K. Kaski, J. Kertész, and A.-L. Barabási, Proc. Nat. Acad. Sci. USA **104**, 7332 (2007).

[11] P. Shu, M. Tang, K. Gong, and Y. Liu, Chaos **22**, 043124 (2012).

[12] M. Karsai, N. Perra, and A. Vespignani, Sci. Rep. **4**, 4001 (2014).

[13] K. Sun, A. Baronchelli, and N. Perra, Eur. Phys. J. B **88**, 1 (2015).

[14] N. Perra, B. Gonçalves, R. Pastor-Satorras, and A. Vespignani, Sci. Rep. **2**, 1 (2012).

[15] S. Liu, N. Perra, M. Karsai, and A. Vespignani, Phys. Rev. Lett. **112**, 1 (2014).

[16] A. Rizzo, M. Frasca, and M. Porfiri, Phys. Rev. E **90**, 1 (2014).

[17] L. Zino, A. Rizzo, and M. Porfiri, Phys. Rev. Lett. **117**, 1 (2016).

[18] G. Petri and A. Barrat, Phys. Rev. Lett. **121**, 228301 (2018).

[19] L. Zino, A. Rizzo, and M. Porfiri, SIAM J. Appl. Dyn. Syst. **17**, 2830 (2018).

[20] A. Rizzo and M. Porfiri, Eur. Phys. J. B **89**, 20 (2016).

[21] D. Li, D. Han, J. Ma, M. Sun, L. Tian, T. Khouw, and H. E. Stanley, EPL **120**, 28002 (2017).

[22] M. Starnini and R. Pastor-Satorras, Phys. Rev. E **89**, 1 (2014).

[23] C. Bongiorno, L. Zino, and A. Rizzo, in *Proc. 57th IEEE Conf. Dec. Control* (2019) pp. 6210–6215.

[24] C. Bongiorno, L. Zino, and A. Rizzo, Appl. Net. Sci. **4**, 1 (2019).

[25] Y. Lei, X. Jiang, Q. Guo, Y. Ma, M. Li, and Z. Zheng, Phys. Rev. E **93**, 1 (2016).

[26] M. Nadini, A. Rizzo, and M. Porfiri, IEEE Trans. Net. Sci. Eng. (2018), published online.

[27] H. Kim, M. Ha, and H. Jeong, Eur. Phys. J. B **88**, 315 (2015).

[28] H. Kim, M. Ha, and H. Jeong, Phys. Rev. E **97**, 062148 (2018).

[29] A. Rizzo, B. Pedalino, and M. Porfiri, J. Theor. Biol. **394**, 212 (2016).

[30] M. J. Keeling and P. Rohani, Ecol. Lett. **5**, 20 (2002).

[31] L. Lü, C.-H. Jin, and T. Zhou, Phys. Rev. E **80**, 046122 (2009).

[32] L. Lü and T. Zhou, Physica A **390**, 1150 (2011).

[33] S. G. Shandilya and M. Timme, New J. Phys. **13**, 13004 (2011).

[34] M. T. Angulo, J. A. Moreno, G. Lippner, A.-L. Barabsi, and Y.-Y. Liu, J. Royal Soc. Interface **14**, 20160966 (2017).

[35] H. Liao and A. Zeng, Sci. Rep. **5**, 11404 (2015).

[36] M. Porfiri and M. R. Marin, IEEE Trans. Net. Sci. Eng. **5**, 42 (2018).

[37] A. Braunstein, A. Ingrosso, and A. P. Muntoni, J. Royal Soc. Interface **16**, 20180844 (2019).

[38] C. Ma, H.-S. Chen, Y.-C. Lai, and H.-F. Zhang, Phys. Rev. E **97**, 022301 (2018).

[39] X. Han, Z. Shen, W.-X. Wang, and Z. Di, Phys. Rev. Lett. **114**, 028701 (2015).

[40] B. Prasse and P. Van Mieghem, IEEE Trans. Net. Sci. Eng. (2018), published online.

[41] T. P. Peixoto, arXiv preprint: 1903.10833 (2019).

[42] M. Tumminello, S. Micciche, F. Lillo, J. Piilo, and R. N. Mantegna, PLOS ONE **6**, e17994 (2011), 1008.1414.

[43] C. Bongiorno, A. London, S. Miccichè, and R. N. Mantegna, Phys. Rev. E **96**, 1 (2017).

[44] J. Fournet and A. Barrat, PLOS ONE **9** (2014).

[45] B. Gonçalves, N. Perra, and A. Vespignani, PLOS ONE **6**, e22656 (2011).

[46] Similar to [24], the assumption $d_i \geq 1$, for all $i \in V$, can be removed with a slight modification of (1).

[47] A matrix is said to be row-stochastic if it nonnegative (entrywise) and each row sums to 1.

[48] N. T. J. Bailey, *The Mathematical Theory of Infectious Diseases and its Applications*, 2nd ed. (Griffin, London, UK, 1975).

[49] In principle for an arbitrary network model, this quantity might also attain negative values.

[50] Y. Benjamini and Y. Hochberg, J. Royal Stat. Soc. B , 289 (1995).

[51] M. E. J. Newman, Phys. Rev. E **67**, 13 (2003).

[52] R. Pastor-Satorras, A. Vázquez, and A. Vespignani, Phys. Rev. Lett. **87**, 258701 (2001).

[53] W. Aiello, F. Chung, and L. Lu, Exp. Math. **10**, 53 (2001).

[54] L. Page, S. Brin, R. Motwani, and T. Winograd, *The PageRank Citation Ranking: Bringing Order to the Web*, Technical Report (1998).

[55] SocioPatterns, "SocioPatterns Primary School Temporal Network Data," `https://www.sociopatterns.org/datasets/primary-school-temporal-network-data`, online; accessed July 12, 2019.