Applied Network Science

RESEARCH Open Access

A novel framework for community modeling and characterization in directed temporal networks

Christian Bongiorno[1†], Lorenzo Zino[2†] and Alessandro Rizzo[1,3*]

*Correspondence:
alessandro.rizzo@polito.it
†Christian Bongiorno and Lorenzo Zino contributed equally to this work.
[1]Department of Electronics and Telecommunications, Politecnico di Torino, Corso Duca degli Abruzzi, 24, 10129 Torino, Italy
[3]Office of Innovation, New York University Tandon School of Engineering, 11201 Brooklyn NY, US
Full list of author information is available at the end of the article

Abstract
We deal with the problem of modeling and characterizing the community structure of complex systems. First, we propose a mathematical model for directed temporal networks based on the paradigm of activity driven networks. Many features of real-world systems are encapsulated in our model, such as hierarchical and overlapping community structures, heterogeneous attitude of nodes in behaving as sources or drains for connections, and the existence of a backbone of links that model dyadic relationships between nodes. Second, we develop a method for parameter identification of temporal networks based on the analysis of the integrated network of connections. Starting from any existing community detection algorithm, our method enriches the obtained solution by providing an in-depth characterization of the very nature of the role of nodes and communities in generating the temporal link structure. The proposed modeling and characterization framework is validated on three synthetic benchmarks and two real-world case studies.

Keywords: Activity driven network, Backbone, Community structure, Heterogeneity, Parameter identification, Time-varying network

Introduction

Many seminal studies have revealed that communities are ubiquitous in networked systems of diverse nature. In fact, community structures have been identified in social, financial, biological, and in many other networks (Girvan and Newman 2002; Newman 2006; Estrada 2011; Benson et al. 2016; Yang and Leskovec 2014). Such communities typically have a complex structure: they present a hierarchical and overlapping organization (Ravasz et al. 2002; Palla et al. 2005; Lancichinetti et al. 2009; Pons and Latapy 2011) and their components have heterogeneous attitudes in the link formation process (Palla et al. 2007), acting as *sources*, mainly generating connections, or as *drains*, on the contrary. Further challenges have to be tackled toward a comprehensive analysis of real-world systems. First, evidence suggests that the patterns of connections between the components of a real-world networked system evolve in time (Volz and Meyers 2008; Holme and Saramäki 2012; Pastor-Satorras et al. 2015; Latapy et al. 2018). Second, the nature of interactions in many biological and technological systems has an inherent direction and is thus nonsymmetrical (Leicht and Newman 2008). Finally, there exist connections that are generated by dyadic relationships between nodes, rather than by properties of the single node. These connections give rise to a link structure, which is often referred as the *irreducible backbone* of temporal interactions, or the structure of *strong ties* (Onnela et al. 2007; Gemmetto et al. 2017). These factors often challenge the applicability of existing temporal network models and community detection algorithms (Lancichinetti et al. 2011; Fortunato and Hric 2016; Khan and Niazi 2017; Schaub et al. 2017; Zhang et al. 2018).

In this work, we propose a novel and extremely flexible model of temporal networks that encompasses many complex features of real-world systems. We refer to this model as *routed activity driven networks* (rADN). It extends the paradigm of activity driven networks (ADNs), which have emerged as a valuable framework to represent and study time-varying networks of interactions (Perra et al. 2012). The main strength of ADNs lays in their simplicity: the time-varying nature of the network is indeed encapsulated in a single parameter vector, called activity, which quantify the propensity of each node to generate transitory connections with the others. Such an activity parameter vector can be easily inferred from empirical data (Perra et al. 2012; Karsai et al. 2014; Liu et al. 2014; Rizzo et al. 2016). Since their original formulation in (Perra et al. 2012), many features have been included into the ADN paradigm toward realistic modeling of complex networks of interactions. These features include a continuous-time formulation of the framework (Zino et al. 2016; 2017), the heterogeneous propensity of nodes to receive connections (Pozzana et al. 2017; Alessandretti et al. 2017), memory mechanisms in the link generation process (Karsai et al. 2014; Sun et al. 2015; Zino et al. 2018), the partition of nodes into a simple community structure (Nadini et al. 2018b), and the presence of an irreducible backbone of recurrent connections (Lei et al. 2016; Nadini et al. 2018a). The simple formulation of ADNs and their extensions are amenable to analytical treatment, allowing for the study of many phenomena on time-varying heterogeneous networks, including epidemic outbreaks (Rizzo et al. 2014, 2016; Petri and Barrat 2018), diffusion of innovation (Rizzo and Porfiri 2016), opinion dynamics (Li et al. 2017), and percolation problems (Starnini and Pastor-Satorras 2014).

Our model incorporates the following features: *i*) the heterogeneity in the propensity to form links with other network nodes; *ii*) the directionality of such links; *iii*) a hierarchical and overlapping time-invariant community structure; *iv*) the heterogeneous involvement of nodes within their communities, acting as sources or drains; and *v*) the presence of an irreducible backbone. The model relies on a relatively compact parameter set able to elicit the complexity of the system in an elegant and intelligible form, highlighting the community structure and its role in the network formation process.

The main goal of this work is to provide a comprehensive and effective means to describe a complex and heterogeneous system through a mesoscopic characterization at the community level. We believe that such a characterization is of great interest for several applications (e.g., epidemic containment, contrast of misinformation) for which, on the one hand, macroscopic mean-field approaches fail in accounting for the inherent diversity throughout the system, whereas the microscopic characterization at the node level comprises an extremely large parameter set, hampering analytical tractability and making time- and resource-consuming Monte Carlo simulations the only way to shed light on the system properties. A preliminary version of this model was presented in (Bongiorno et al. 2018). Here, we extend such a preliminary work by adding the irreducible backbone of strong ties, a more flexible and comprehensive connection mechanism, and a more detailed analysis and assessment of the proposed model over synthetic benchmarks and two real-world datasets.

The identification of the model parameters from temporal link formation data poses a series of challenges, since the occurrence of a link cannot be unequivocally attributed the mechanism that has generated it. Hence, differently from existing methods for community detection that tend to explain link formation as a sole consequence of the community structure, here we establish a probabilistic framework to quantify the belief in alternative link formation processes, accounting for three co-existing mechanisms of connection: *i)* communities, *ii)* community-free, and *iii)* backbone.

The proposed parameter identification and community detection strategies unfold over three main steps, starting from the observation of the links generated during a given time-window, where it is assumed that both the community structure and the backbone do not change in time. Using an integrated version of the observed temporal network of contacts, we apply an existing community detection method (Fortunato and Hric 2016; Khan and Niazi 2017) to infer a reference community structure for our rADN-based model. Then, we perform parameter identification of the rADN model solving a quadratic optimization program with linear constraints (Boyd and Vandenberghe 2004). Such an identification problem is naturally underdetermined, implying that a family of rADN models with the same community structure but different parameters may equivalently reproduce the link formation process. In fact, different parameter combinations lead to different probabilistic explanations of the link formation, relying on different probabilistic blends of the three mechanisms mentioned above. A free parameter vector, called *community belief*, is thus defined to quantify the belief in the role of communities in the network formation: when the components of the community belief approach one, we tend to assume that the connection mechanism is mostly governed by communities, otherwise, when they approach zero, the role of communities becomes negligible. Hence, a family of models may be obtained by running an identification procedure for each value of the community belief. Confidence intervals on the community belief can then be established, in order to obtain a family of models that is practically compatible with the available data. Thanks to the belief mechanism, the use of our method in conjunction with different preliminary community detection algorithms allows us to quantitatively compare different hypotheses on the community structure and the link formation process.

We successfully validate our approach on three synthetic benchmarks that exhibit different features, all generated through the proposed rADN model. We then apply our method to two different real-world case studies. The first is based on the Enron email corpus (Cohen), where no information is available of the community structure. Here, we propose our method as a tool to compare and assess the outputs of different community detection algorithms. The second case study uses data about face-to-face interactions in a primary school (SocioPatterns). Here, we use metadata on the partition of students in classes to provide a ground truth on the community structure. In this case, our method is used to improve the characterization of the community structure.

The rest of this paper is organized as follows. In "Model" section, we formalize the rADN model. In "Estimation of the model parameters from empirical data" section, we propose a method to estimate the parameters of a rADN from empirical data. In "Validation on synthetic networks" section, we validate the proposed method on three synthetic networks, exhibiting different features of real-word networks such as heterogeneity, the presence of a backbone, overlapping and hierarchical communities. "Case studies" section is devoted to the analysis of the two case studies. Finally, "Conclusion" section concludes the paper and outlines our future research.

Model

A rADN is a network composed by a set of n nodes $\mathcal{V} = \{1, \dots, n\}$ connected through a time-varying link structure $\mathcal{G}(t) = (\mathcal{V}, \mathcal{E}(t))$, where $\mathcal{E}(t) \subseteq \mathcal{V} \times \mathcal{V}$ denotes the time-varying link set. Links are generated according to a continuous-time mechanism, following the formalism proposed in (Zino et al. 2016; 2017). The continuous-time formulation allows for addressing some theoretical limitations posed by the original discrete-time formulation of ADNs (Perra et al. 2012) and is not subject to the issues related to the choice of the discrete time step (Ribeiro et al. 2013).

A positive (time-invariant) *activity rate* $a_i > 0$ is assigned to each node $i \in \mathcal{V}$. The activity rate quantifies the node's propensity to generate transitory connections with other nodes in the network, as detailed in the following. Activity rates are gathered in a n-dimensional vector a. We define the *routing matrix* $P \in [0, 1]^{n \times n}$, as a stochastic matrix (i.e., an entry-wise nonnegative matrix such that all rows sum to one) with zero diagonal entries. The entry P_{ij} of the routing matrix measures the propensity of node i to generate connections toward node j, as detailed in the following.

Hence, the triple $(\mathcal{V}, a, P)$ identifies a rADN. Connections are generated according to a similar mechanism to the one of standard continuous-time ADNs (Zino et al. 2016), except for a nonuniform choice of the connection wirings, which are governed by the routing matrix P, whose construction will be detailed later in this section. The following algorithm summarizes the evolution of a rADN:

1 at $t = 0$, the link set is set as empty ($\mathcal{E}(0) = \emptyset$) and a Poisson clock (Bailey 1990) with rate a_i (each one independent of the others) is initialized for each node $i \in \mathcal{V}$;
2 if at time t the clock associated with node i clicks, then node i activates and randomly selects a node $j \in \mathcal{V}$ to connect to with probability P_{ij};
3 the directed link (i, j) is instantaneously added to the link set $\mathcal{E}(t)$; and
4 link (i, j) is immediately removed from the link set, the Poisson clock associated with node i is re-initialized, and the algorithm is resumed to item 2.

In their original formulation, links of continuous-time ADNs are ephemeral. Even though this could seem an over simplification, many interactions in social and biological systems have a negligible duration with respect to the time scale of the network evolution and of the emerging phenomena of the system. Relevant examples are e-mails or messages exchanged in social networks or physical interactions between individuals. The model can be straightforwardly extended by including nonephemeral connections, by modifying item 4 of the previous algorithm. For instance, one could remove link (i, j) after a certain time-interval, which could be fixed or drawn at random from any distribution.

According to our mechanism, the occurrences of the directed link (i, j) are governed by a (split) Poisson process (Ross 2009), whose rate is equal to

$$A_{ij} = a_i P_{ij}. \tag{1}$$

The link activation rates A_{ij} can be gathered into the *activity rate matrix* $A \in \mathbb{R}_+^{n \times n}$, which encapsulates the information both on the activity rate vector a, and on the routing matrix

P. Therefore, a rADN is completely identified by the couple $(\mathcal{V}, A)$. Given A, the activity rate vector a and the routing matrix P can be retrieved as

$$a_i = \sum_{j \in \mathcal{V}} A_{ij}, \qquad P_{ij} = \frac{A_{ij}}{\sum_{h \in V} A_{ih}}, \qquad i,j \in \mathcal{V}. \tag{2}$$

Fixing a time-window of duration $T > 0$, we introduce the *weighted integrated network* over the time-window, represented by the pair $\mathcal{G}_T = (\mathcal{V}, W)$, where $\mathcal{V}$ is the node set, which coincides with the node set of the time-varying network, and $W \in \mathbb{Z}_+^{n \times n}$ is a *weighted adjacency matrix*. Specifically, W_{ij} counts the number of occurrences of directed links from node i to node j in the time-window of duration T. In rADNs, the entry W_{ij} of the weighted adjacency matrix is a Poisson distributed random variable with parameter $A_{ij}T$ (Ross 2009). Hence, the probability of observing w occurrences of the link (i,j) in the integrated network $\mathcal{G}_T$ is equal to

$$\mathbb{P}[\, W_{ij} = w] = \frac{\left(A_{ij}T\right)^{w}}{w!} \exp\{-A_{ij}T\}. \tag{3}$$

Figure 1 illustrates an example of the construction of an integrated network from link observations over a given time-window of duration $T = 1$. In this example, the weighted adjacency matrix is equal to

$$W = \begin{bmatrix} 0 & 2 & 2 \\ 1 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix}. \tag{4}$$

Routing matrix *P*

The routing matrix P is constructed to encapsulate the information on *i*) the organization of nodes in nontrivial, hierarchical and overlapping communities; *ii*) the heterogeneous involvement of nodes in their communities, characterized through the presence of sources and drains; and *iii*) the existence of an irreducible backbone.

In order to distinguish the contribute of the community structure from that of the backbone to the link generation process, we define matrix P as a convex combination of two $n \times n$ stochastic matrices C and R. The convex combination is weighted by a n-dimensional nonnegative (entry-wise) vector $\lambda \in [0, 1]^n$, as

$$P_{ij} = \lambda_i C_{ij} + (1 - \lambda_i) R_{ij}, \qquad i,j \in \mathcal{V}, \tag{5}$$

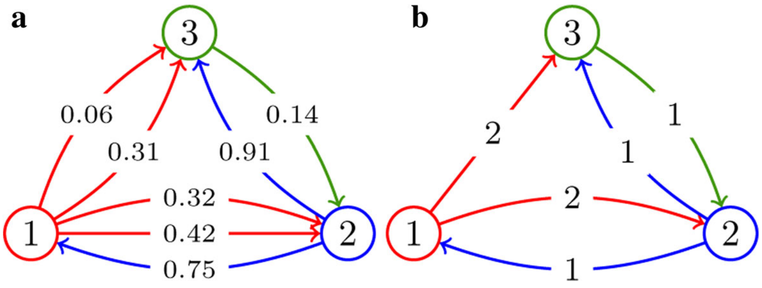

Fig. 1 Exemplification of the construction of an integrated network in a time-window of duration $T = 1$. In (**a**), occurrences of links are plotted separately, along with their time-stamps (2 decimal digits). In (**b**), the integrated network $\mathcal{G}_T$ is illustrated. Weights represent the number of occurrences of each link in the time-window

where matrix C, named *community matrix*, encodes the information on the role of the community-based mechanism in the link generation process, while matrix R encodes the role of the backbone in the process and is called *backbone matrix*. Toward a compact formalization of the model, the community-free mechanism is considered as a special case of the community-based mechanism, through the inclusion into matrix C of the special *all-to-all community*, as detailed in "Community matrix C" section. In the following, unless specified differently, the term *community-based* will refer to both the community-based and the community-free mechanisms, thus leaving aside only the backbone mechanism. Specifically, entry C_{ij} is the probability that node i connects to j as a consequence of the community structure (including the all-to-all community), while entry R_{ij} is the probability that such a connection is generated as a consequence of a dyadic relationship (backbone) between i and j. The entry λ_i weights the two mechanisms by quantifying the strength of the community-based mechanism in the process of link formation from node i. In general, the entries of vector λ are nonuniform, to capture the heterogeneous influence of the two mechanisms for different nodes. The limit case $\lambda_i = 0$ represents the scenario in which the community structure has no influence on the link generation process from node i and all its links are caused by the presence of the irreducible backbone, while the case $\lambda_i = 1$ models the case in which node i wires its connections only driven by the community structure.

The mechanism governed by the convex combination in (5), illustrated in Fig. 2, has an immediate probabilistic interpretation. When node i activates, connections are driven by the community structure with probability λ_i, whereas they are driven by the backbone mechanism with probability $1 - \lambda_i$. In the following, we will detail the construction principles of matrices C and R.

Community matrix C

Here, matrix C is designed to model a time-invariant community structure. This simplifying assumption is reflected in many real-world systems, where the pace of evolution of the community structure is much slower than the link generation process, as in (Bao and

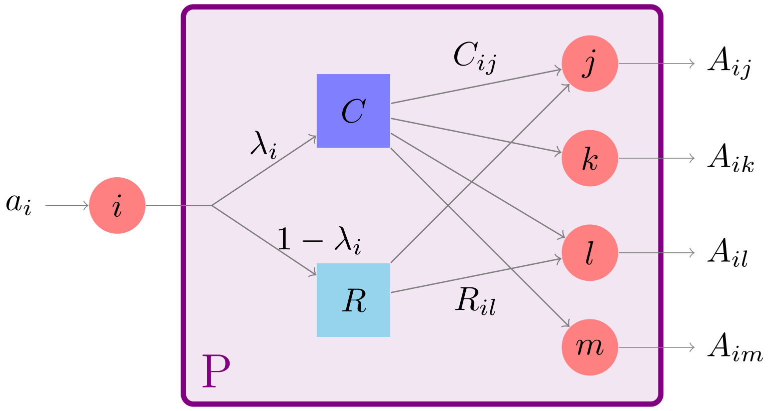

Fig. 2 Schematic of the mechanism governing a rADN model. Node $i \in \mathcal{V}$ activates with rate a_i. Then, with probability λ_i it generates a connection due the community-based mechanism, i.e., following the probabilities in the community matrix C. Otherwise, with probability $1 - \lambda_i$, the link is caused by the backbone in matrix R. These mechanisms yield the corresponding link activation rates in matrix A

Michailidis 2018). Different scenarios, where the community structure evolves in time, can be found in Rossetti and Cazabet (2018).

Given a time-invariant set of $k \geq 0$ nontrivial communities, we label them with positive integer numbers $h \in \{1, \ldots, k\}$. Trivial communities are the empty set, singletons, and the whole node set $\mathcal{V}$. To model the community-free mechanism, we add an *all-to-all* trivial community that coincides with the whole system $\mathcal{V}$, labeled by index 0. Hence, the *community set* $\mathcal{K} = \{0, \ldots, k\}$ comprises the trivial *all-to-all* community 0 and k nontrivial communities. Considering the hth community, we denote by $\mathcal{V}_h \subseteq \mathcal{V}$ the set of nodes that belong to it, while $n_h := |\mathcal{V}_h|$ is its cardinality. On the other hand, considering the generic node $i \in \mathcal{V}$, we denote by $\mathcal{C}_i := \{h : i \in \mathcal{V}_h\}$ the set of communities to which node i belongs, and with $c_i := |\mathcal{C}_i|$ its cardinality.

We observe that the rADN paradigm allows each network node to belong to an arbitrary number of communities. This encompasses and generalizes the paradigm of modular ADNs (Nadini et al. 2018b), where each node belongs to exactly one nontrivial community. The heterogeneous attitude of nodes in their different communities (Palla et al. 2007) is modeled by defining a stochastic rectangular matrix $Q \in [0, 1]^{n\times(k+1)}$, named *community strength matrix*, such that $Q_{ih} > 0$ if and only if $i \in \mathcal{V}_h$. The entry Q_{ih} is the probability that a link from node i that is caused by a community-based mechanism is wired within the hth community. The entries of matrix Q can be thus interpreted as the importance that each node gives to the each of the communities it belongs to. If Q_{ih} is large, node i acts as a source in community h, generating many inter-community links. On the other hand, if Q_{ih} is small, then node i will act as a drain, mostly receiving connections from other members. In this perspective, matrix Q quantifies the strength of the *active involvement* of nodes in each of their communities.

Formally, we define the community matrix C, entry-wise, as

$$C_{ij} = \begin{cases} 0 & \text{if } i = j \\ \sum\limits_{h \in \mathcal{C}_j} Q_{ih} \frac{1}{n_h - 1} & \text{otherwise.} \end{cases} \tag{6}$$

Figure 3 illustrates the mechanisms that govern the formation of the community matrix C. Specifically, when node i generates a connection following the community-based

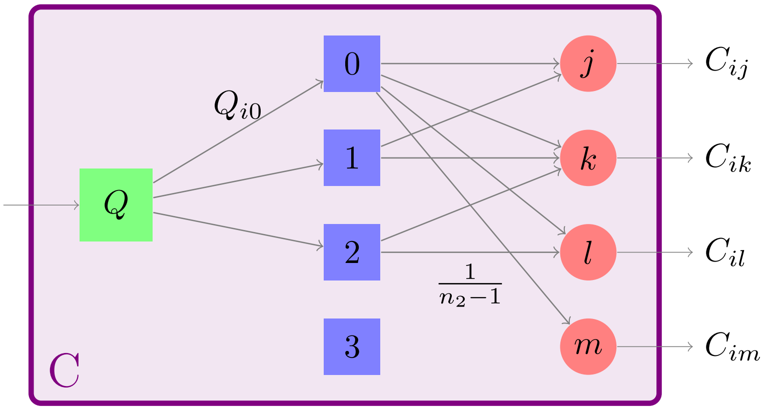

Fig. 3 Schematic of the mechanism that governs the community matrix C. First, a community is selected, according to the probabilities in the community strength matrix Q. Then, a node in the selected community is chosen uniformly at random

mechanism, first, it randomly selects a community $h \in \mathcal{K}$ to which it belongs, according to the probabilities in the ith row of matrix Q. Then, it connects to a node j chosen uniformly at random among the $n_h - 1$ nodes of the hth community (excluding node i).

Backbone matrix R

Similar to the community structure, also the irreducible backbone is typically fixed in time or it evolves much slower than the link formation process (Onnela et al. 2007; Gemmetto et al. 2017). Hence, here we hypothesize that this is constant for time-windows of reasonable duration.

The backbone is thus modeled by a time-invariant graph $G_R = (\mathcal{V}, \mathcal{E}_R)$ and by a stochastic matrix $R \in [0, 1]^{n \times n}$, such that $R_{ij} > 0$ if and only if $(i, j) \in \mathcal{E}_R$. The entry R_{ij} measures the strength of the dyadic relationship between node i and node j in a probabilistic framework. In many real-world scenarios of social and biological systems, it is reasonable to assume matrix R to be sparse, as observed from many empirical data sources (Newman 2003; Ballerini et al. 2008).

We remark that, in the limit case where node i is not influenced by the backbone mechanism, i.e., $\lambda_i = 1$, the entries of the ith row of the backbone matrix R have no influence on the link formation mechanism. Without any loss in generality, in this case we set the corresponding rows of matrix R equal to the corresponding rows of a $n \times n$ identity matrix, i.e., we set $R_{ii} = 1$ and $R_{ij} = 0$, for any $j \neq i$.

The set of parameters that characterize a rADN is summarized in Table 1. To recapitulate, taking into account the construction mechanism of matrix P, we detail the algorithm that summarizes the evolution of a rADN as follows:

1 at $t = 0$, the link set is set as empty ($\mathcal{E}(0) = \emptyset$) and a Poisson clock (Bailey 1990) with rate a_i (each one independent of the others) is initialized for each node $i \in \mathcal{V}$;

2 if at time t the clock associated with node i clicks, then node i activates and randomly chooses whether the connection is generated a) by the community-based mechanism (with probability λ_i), or b) by the backbone (with probability $1 - \lambda_i$). Then, depending on the previous choice, either a) or b) occurs, where

   a) node i randomly selects a community $h \in \mathcal{K}$. Specifically, community h is selected with probability Q_{ih}. Then, node j is chosen uniformly at random among the $n_h - 1$ nodes of the hth community excluding node i; or

Table 1 Parameters that characterize a rADN model

n	Number of nodes
k	Number of (nontrivial) communities
$\mathcal{K}$	Set of communities
$\mathcal{V}_h$	Set of nodes in the hth community
n_h	Number of nodes in the hth community
$\mathcal{C}_i$	Set of communities to which node i belongs
c_i	Number of communities to which node i belongs
a	Activity rate vector
P	Routing matrix
λ	Community weight vector
C	Community matrix
R	Backbone matrix
Q	Community strength matrix

   b) node i randomly selects a node $j \in \mathcal{V}$. Specifically, node j is selected with probability R_{ij};

3 the directed link (i,j) is instantaneously added to the link set $\mathcal{E}(t)$; and

4 link (i,j) is immediately removed from the link set, the Poisson clock associated with node i is re-initialized, and the algorithm is resumed to item 2.

We observe that our extended rADN modeling framework actually encompasses many variants of standard ADNs proposed in the recent literature. Some relevant examples are presented in the following.

Standard ADNs

Standard continuous-time ADNs (Zino et al. 2016) are obtained by setting $\lambda_i = 1, i \in \mathcal{V}$, and $\mathcal{K} = \{0\}$. This choice yields $P_{ii} = 0$, for any $i \in \mathcal{V}$, and

$$P_{ij} = C_{ij} = \frac{Q_{i0}}{n-1} = \frac{1}{n-1}, \qquad i \in \mathcal{V}, j \in \mathcal{V} \setminus \{i\}. \tag{7}$$

Modular ADNs

Modular ADNs (Nadini et al. 2018b) can be derived as a particular case of rADN with $\lambda_i = 1$, $\forall\, i \in \mathcal{V}$. Here, the set of communities $\mathcal{K}$ defines a partition of the node set where each node belongs to one and only one community. The notation used in the original incarnation defined in Nadini et al. (2018b) can be retrieved by setting $Q_{i0} = 1-\mu$, $Q_{ih} = \mu, \forall\, i \in \mathcal{V}_h$. In this case, (6) reads $P_{ii} = 0, \forall\, i \in \mathcal{V}$, and, for any $j \neq i$,

$$P_{ij} = C_{ij} = \begin{cases} \frac{1-\mu}{n-1} + \frac{\mu}{n_h-1} & \text{if } i \in \mathcal{V}_h,\ j \in \mathcal{V}_h, \\ \frac{1-\mu}{n-1} & \text{if } i \in \mathcal{V}_h,\ j \notin \mathcal{V}_h. \end{cases} \tag{8}$$

ADNs with attractiveness

Attractiveness has been added to ADNs to model the heterogeneous propensity of nodes to receive connections (Pozzana et al. 2017; Alessandretti et al. 2017). Specifically, given an attractivity vector $b > 0$ (entry-wise), the probability that a node generates a link to node $j \in \mathcal{V}$ is proportional to b_j. In the framework of rADNs, this feature can be modeled by setting $\lambda_i = 0$, $\forall\, i \in \mathcal{V}$, and all the entries of the backbone matrix R as

$$R_{ij} = \frac{b_j}{\sum\limits_{k \in \mathcal{V} \setminus \{i\}} b_k}, \qquad \forall\, i,j \in \mathcal{V}. \tag{9}$$

Estimation of the model parameters from empirical data

Here, we develop a technique to identify the model parameters. Specifically, we estimate the activity rate vector a, the weight vector λ, the community strength matrix Q, and the backbone matrix R. The objective of the technique presented in this section is to devise a procedure to deepen the characterization of communities, shading light on the diverse role of their members and their role in the link formation process.

Parameter identification procedure

As stated in the introduction, we preliminary use an existing community detection algorithm, feeding it with the temporal sequence of the link formation, and obtaining as output *i*) the realization of an integrated version of the temporal network over a time-window of duration T, whose information is gathered in the weighted adjacency matrix

W; and *ii*) the community set $\mathcal{K}$ and the sets $\mathcal{V}_h$, $h \in \mathcal{K}$, obtained as the output of the community detection algorithm. According to (Perra et al. 2012), the activity rate vector can be estimated as

$$\hat{a}_i = \sum_{j \in \mathcal{V}} \frac{W_{ij}}{T}, \qquad i \in \mathcal{V}. \tag{10}$$

The expected number of occurrences of the link (i,j), denoted by $\bar{W}_{ij}$, is computed following (3), and it is equal to

$$\bar{W}_{ij} = \hat{a}_i T \sum_{h \in \mathcal{C}_i} \lambda_i Q_{ih} \frac{1}{n_h - 1} + \hat{a}_i T(1 - \lambda_i) R_{ij}. \tag{11}$$

In order to estimate the other model parameters, i.e., the community weight vector λ, the community strength matrix Q, and the backbone matrix R, we formulate the identification problem in terms of a constrained optimization program. We observe that the identification problem is naturally underdetermined. In fact, the observed data consists of a $n \times n$ matrix, while the set of parameters to be estimated comprises a $n \times n$ matrix, a $n \times (k + 1)$ matrix, and a n-dimensional vector. Hence, except for unlikely particular cases, the number of parameters to be estimated exceeds the number of equations that can be written using the available data. To address this issue, we introduce a free parameter vector $\gamma \in [0, 1]^n$, named *community belief*, which measures our belief in the prominence of the role of the community-based mechanism in the link formation process. Tuning this parameter vector is the most delicate task in the application of our method. In "Confidence interval for the community belief parameter" section, we put forward a statistical procedure to assess a confidence interval for such a parameter vector. In particular, we identify the largest value for γ that is compatible with the observed data, which yields the characterization of the system with the highest belief in the community-based link formation mechanism. Such a model is often preferred, as it leads to a characterization at a mesoscopic level, whereby the system characteristics are captured with a good detail and an intermediate granularity, which ensures a good description of the system behavior without incurring in the issues related to a microscopic, node-based representation. However, in "Validation on synthetic networks" section we show that when the dataset has a small size, smaller values of the parameter γ within the prescribed confidence interval may be more suitable to describe the system without overfitting the community structure.

The identification problem is formalized by writing a set of n disjoint minimization problems, one for each node $i \in \mathcal{V}$. Specifically, for node $i \in \mathcal{V}$, we want to minimize the function

$$f(\varepsilon_{i\bullet}, Q_{i\bullet}, R_{i\bullet}, \lambda_i) = (1 - \gamma_i) \sum_{j=1}^{n} \varepsilon_{ij}^2 + \gamma_i (1 - \lambda_i) R_{ij}, \tag{12}$$

with respect to variable λ_i and the entries of the ith row of matrices ε, Q and R, written in compact form as $\varepsilon_{i\bullet}$, $Q_{i\bullet}$, and $R_{i\bullet}$, respectively. The minimization problem in (12) is subject to several constraints: we require that the number of occurrences of the link (i,j), i.e., W_{ij}, is equal to its expected value $\bar{W}_{ij}$, computed according to (11), up to some natural statistical fluctuation, modeled by the residual ε_{ij}; we also require the matrices Q and R to be stochastic, and the variable λ_i to be nonnegative and not greater than 1. These constraints are gathered as follows:

$$\begin{cases} \hat{a}_i T \sum\limits_{h \in \mathcal{C}_i} \lambda_i Q_{ih} \frac{1}{n_h - 1} + \hat{a}_i T (1 - \lambda_i) R_{ij} + \varepsilon_{ij} = W_{ij} & \forall j \in \{1, \ldots, n\}, \\ \sum\limits_{h \in \mathcal{C}_i} Q_{ih} = 1, & \\ \sum\limits_{j=1}^{n} R_{ij} = 1, & \\ 0 \leq Q_{ih} \leq 1, & \forall h \in \mathcal{C}_i, \\ 0 \leq R_{ij} \leq 1, & \forall j \in \{1, \ldots, n\}, \\ 0 \leq \lambda_i \leq 1. & \end{cases} \tag{13}$$

We observe that the objective function in (12) consists of the sum of two terms: the first summand is the sum of the squared residuals, whose minimization allows for obtaining a model that is compatible with the observed data; the second summand is a cost related to the contributions of the backbone-based mechanism. In the absence of the second term, a trivial solution would be $\lambda_i = 0$ and $R_{ij} = \bar{W}_{ij}/\hat{a}_i$, that is, the whole link formation process is explained in terms of dyadic relationships between nodes. However, this is often not consistent with the empirical observation of a sparse backbone in systems of different nature (Ballerini et al. 2008; Newman 2003), and it fails to provide a description of the system at a mesoscopic scale.

Although the constraints are nonlinear, the change of variable

$$\tilde{Q}_{ih} := \lambda_i Q_{ih} \qquad \tilde{R}_{ij} := (1 - \lambda_i) R_{ij}, \tag{14}$$

allows us to write (12) as a quadratic programming problem (Boyd and Vandenberghe 2004) with linear constraints, which can be solved with a reasonable computational effort. Specifically, the objective function reads

$$f\left(\varepsilon_{i\bullet}, \tilde{Q}_{i\bullet}, \tilde{R}_{i\bullet}, \lambda_i\right) = (1 - \gamma_i) \sum_{j=1}^{n} \varepsilon_{ij}^2 + \gamma_i \tilde{R}_{ij}, \tag{15}$$

subject to the following constraints:

$$\begin{cases} \hat{a}_i T \sum\limits_{h \in \mathcal{C}_i} \tilde{Q}_{ih} \frac{1}{n_h - 1} + \hat{a}_i T \tilde{R}_{ij} + \varepsilon_{ij} = W_{ij} & \forall j \in \{1, \ldots, n\}, \\ \sum\limits_{h \in \mathcal{C}_i} \tilde{Q}_{ih} + \sum\limits_{j=1}^{n} \tilde{R}_{ij} = 1, & \\ 0 \leq \tilde{Q}_{ih} \leq 1, & \forall h \in \mathcal{C}_i, \\ 0 \leq \tilde{R}_{ij} \leq 1, & \forall j \in \{1, \ldots, n\}. \end{cases} \tag{16}$$

The computational complexity of the parameter identification method proposed here can be estimated as a function of the number n of nodes in $\mathcal{V}$, the number m of nonzero entries of the weighted adjacency matrix W, and the number k of communities. Specifically, we obtain that the computational complexity is equal to $O\left(n^2 + nm + nk\right)$. Since it often holds $k << n$, we conclude that, for sparse integrated networks the computational complexity is $O\left(n^2\right)$, while for dense networks it is $O\left(n^3\right)$.

Confidence interval for the community belief parameter

In this section, we develop a statistical method to identify a range of values for the community belief vector γ, compatible with the natural statistical fluctuations. Since each of the components of the vector γ is derived through a distinct minimization problem, here we focus on the generic ith component, independently of the others. In the previous section, we observed that, for $\gamma_i = 0$, problem (15) admits the trivial solution in which the whole link formation process from node i is explained in terms of the backbone. Then, when γ_i grows, the role of communities in the process of link generation gains more importance. When γ_i is too large, however, the role of communities in the link formation process might be overestimated and the models obtained for these values of the community belief parameter are not statistically compatible with the data observed. Here, we put forward a technique to identify the largest value of the community belief γ_i such that the parameters of the rADN model identified for that value of γ_i are compatible with the available data.

Fixing a value of the parameter γ_i, the minimization problem (15) can be solved by means of a quadratic programming solver. We denote the corresponding solution as $Q^{(\gamma_i)}$, $R^{(\gamma_i)}$, and $\lambda^{(\gamma_i)}$. Using the parameters estimated with this solution, and the expression for the probability of link formation in rADN in (3), we can determine the distribution of the ith row of the weighted adjacency matrix W. We denote such a row as $W_{i\bullet}^{(\gamma_i)}$, to stress its dependence on the choice of the parameter γ_i. Specifically, each row entry $W_{ij}^{(\gamma_i)}$ is a Poisson random variable, independent of the others, with expected value

$$\bar{W}_{ij}^{(\gamma_i)} = \lambda_i^{(\gamma_i)} \sum_{h \in C_i} \frac{Q_{ih}^{(\gamma_i)}}{n_h - 1} + \left(1 - \lambda_i^{(\gamma_i)}\right) R_{ij}^{(\gamma_i)}. \tag{17}$$

Hence, the likelihood that the ith row of the observed weighted adjacency matrix W is a realization of the random variable $W^{(\gamma_i)}$ is given by

$$\mathcal{L}\left(\bar{W}^{(\gamma_i)} | W\right) = \prod_{j=1}^{n} \frac{\left(\bar{W}_{ij}^{(\gamma_i)}\right)^{W_{ij}} e^{-\bar{W}_{ij}^{(\gamma_i)}}}{W_{ij}!}, \tag{18}$$

which is the probability that the realization of the n independent Poisson distributed random variables with expected values computed according to (17) coincide with the ith row of the observed matrix W. Since the product of small probabilities is numerically unstable, it is convenient to test the log-likelihood (Boyd and Vandenberghe 2004) instead of (18), which is

$$\log \mathcal{L}\left(\bar{W}^{(\gamma_i)} | W\right) = \sum_{j=1}^{n} \log \left(\frac{\left(\bar{W}_{ij}^{(\gamma_i)}\right)^{W_{ij}} e^{-\bar{W}_{ij}^{(\gamma_i)}}}{W_{ij}!} \right). \tag{19}$$

Unsurprisingly, the value γ_i that maximizes the log-likelihood function is $\gamma_i = 0$, which yields the scenario where communities have no role in the link formation process and all the connections are explained at the microscopic scale of the irreducible backbone. This scenario is the result of data overfitting, which is a well known problem of the methods based on maximum likelihood (Bishop 2006). Specifically, in this case dyadic relationships are overfitted. We observe that this extreme scenario is not particularly interesting in our framework. In fact, the interest in community-based modeling and community detection algorithms is to describe the system at the mesoscopic level, that is, at a level higher than the individual, microscopic one. This implies that the information encapsulated by a usually high number of microscopic parameters is compressed in a much smaller number of mesoscopic parameters, that is, those related to communities. For this reason, we are interested in models that do include communities, that is, with γ greater than zero. In particular, we are interested in finding the model with the largest value of γ_i that produces an rADN model that is compatible with the available data observed in the weighted adjacency matrix W. To this aim, we perform a Likelihood Ratio (LR) test (Casella and Berger 2002). The LR test determines whether the null hypothesis that the observed ith row of the weighted adjacency matrix $W_{i\bullet}$ is obtained from a vector of independent Poisson variables with expected values $\bar{W}_{ij}^{(\gamma_i)}$ from (17), for $j = 1, \ldots, n$, should be rejected. Specifically, fixing a significance coverage $\alpha \in [0, 1]$, the LR test rejects the null hypothesis if the statistic

$$D := 2\left(\log \mathcal{L}\left(\bar{W}^{(0)}|W\right) - \log \mathcal{L}\left(\bar{W}^{(\gamma_i)}|W\right)\right) \tag{20}$$

is greater than a threshold $q_{1-\alpha}$, where $q_{1-\alpha}$ is the $(1-\alpha)$-quantile of a chi-squared distribution with $n-1$ degrees of freedom (Casella and Berger 2002). We remark that the reduction of the degrees of freedom of the distribution from n to $n-1$ is due to the absence of self-loops.

It is worth noticing that the statistic D is a monotonic increasing function of γ_i. Hence, fixing a significance coverage $\alpha \in [0, 1]$, the LR test ultimately identifies a threshold $\bar{\gamma}_i$, equal to the value for which the statistic D is equal to the $(1-\alpha)$-quantile of a chi-squared distribution with $n-1$ degrees of freedom. All the values of $\gamma_i > \bar{\gamma}_i$ are rejected. Hence, our procedure establishes a confidence interval for the parameter γ_i of the form $\gamma_i \in [0, \bar{\gamma}_i]$. Unfortunately, the value $\bar{\gamma}_i$ that produces the significance coverage α cannot be derived analytically. However, since the statistic D is monotone in γ_i, several numerical methods can efficiently retrieve a good approximation of the threshold $\bar{\gamma}_i$ (Hammings 1973).

The technique described above identifies a range for the ith component γ_i of the community belief vector γ that is compatible with the empirical data. Implementing this procedure for all the nodes $i \in \mathcal{V}$, we obtain a confidence interval for the whole parameter vector γ, which is the n-dimensional hyperrectangle $\gamma \in [0, \bar{\gamma}_1] \times \cdots \times [0, \bar{\gamma}_n]$. This yields a set of mathematical models that are compatible with the observed data. The parameter vector may be tuned within this hyperrectangle, depending on the user's belief in the community structure, on the amount of data available (as we will discuss in "Validation on synthetic networks" section) and, possibly, on additional information available on the systems such as historical data, or measurements on similar systems. We refer to the model obtained with $\gamma_i = \bar{\gamma}_i$, for all $i \in \mathcal{V}$, as the model with the largest belief in the community structure, among those compatible with the observed data. The flow chart in Fig. 4 summarizes the whole procedure of our parameter identification method, from the data consisting of a sequence of temporal links, to the definition of an rADN model.

Validation on synthetic networks

In this section, we validate our procedure over three different benchmarks of temporal networks. In the first benchmark, described in "Exclusive heterogeneous communities" section, nodes are partitioned into six exclusive communities of different size, so that each node belongs to exactly one community (and to the trivial, all-to-all one). A time-invariant, irreducible backbone is also present. In the proposed benchmark, nodes present a high level of heterogeneity, both in their global activity and in their involvement in their community. The analysis of this benchmark suggests that our parameter identification method is able to identify the model parameters in presence of heterogeneity in the network structure. In "Hierarchical communities" section, we propose a second benchmark where communities present a hierarchical structure. The third benchmark,

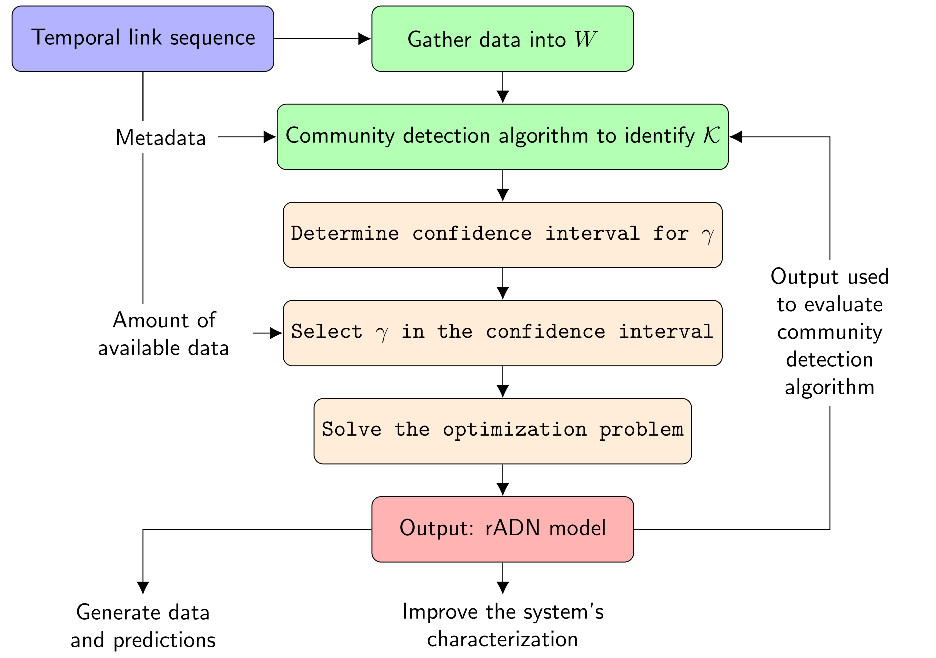
Temporal link sequence
Gather data into W
Metadata
Community detection algorithm to identify $\mathcal{K}$
Determine confidence interval for γ
Amount of available data
Select γ in the confidence interval
Output used to evaluate community detection algorithm
Solve the optimization problem
Output: rADN model
Generate data and predictions
Improve the system's characterization

Fig. 4 Flow chart of the parameter identification method. Data on the temporal connections are gathered into the weighted adjacency matrix W of the integrated network. Then, a community detection algorithm is used to preliminary detect a set of communities. The community detection algorithm may be enriched by available metadata. A confidence interval for the community belief is then computed and the parameter γ is selected within this interval, on the basis the available data. Finally, fixed γ, the model parameters are identified through the solution of an optimization problem. A rADN model statistically compatible with the available data is eventually derived. The output of such a model can be used to assess and improve the performance of existing community detection algorithms or, on its own, to enrich the characterization of the system and to generate data and predictions that are compatible with the available data

characterized by overlapping communities, is discussed in "Overlapping communities" section. Also in these cases, we successfully perform the parameter identification by means of the method proposed in this work.

Exclusive heterogeneous communities

We generate a network with an exclusive community structure, where each node belongs to one and only one nontrivial community. We partition $n = 100$ nodes into $k = 6$ nontrivial communities with heterogeneous size, as shown in Table 2. Thus, the community set $\mathcal{K}$ and the sets $\mathcal{V}_h, h \in \mathcal{K}$ are known.

The community strength matrix Q is randomly constructed, in order to model heterogeneity in the nodes' attitude to generate inter-community links. Specifically, the entries of the first column of Q, which represent the nodes' involvement in the all-to-all community, are selected from independent beta distributions with mean 0.25 and variance 0.02. We remark that the selection of the entries from a beta distribution ensures $Q_{0i} \in [0, 1]$, for any $i \in \mathcal{V}$ (Ross 2009). Matrix Q is fully determined by its first column. In fact, each row of the matrix Q has only two nonzero entries (since each node belongs to a unique nontrivial community) and the matrix is stochastic. Hence, the other nonzero entry of the generic ith row is equal to $1 - Q_{i0}$. Matrix Q obtained according to this procedure is illustrated in Fig. 5a.

Table 2 Benchmark with exclusive community structures

# community	Size n_h	Members $\mathcal{V}_h$
1	50	$\{1,\ldots,50\}$
2	20	$\{51,\ldots,70\}$
3	10	$\{71,\ldots,81\}$
4	10	$\{81,\ldots,90\}$
5	5	$\{91,\ldots,95\}$
6	5	$\{96,\ldots,100\}$

The backbone matrix R is defined as follows. First, we construct the graph $G_R = (\mathcal{V}, \mathcal{E}_R)$, corresponding to the backbone, according to an Erdös-Rényi random graph model (Erdős and Rényi 1959) with parameter $p = 4/99$. Such a choice of the parameter p produces a network with average degree equal to 4. Specifically, link $(i,j) \in \mathcal{E}_R$ with probability p, each link independently of the others. Then, the entries R_{ij}, for $(i,j) \in \mathcal{E}_R$, are assigned uniformly at random, such that each row sums 1, while all the other entries of the row are set to 0. In the extreme case in which node $i \in \mathcal{V}$ has no outgoing links, then we set the diagonal entry $R_{ii} = 1$, all other entries equal to 0, and the corresponding $\lambda_i = 1$. The other entries of the vector λ (i.e., those corresponding to nodes i that have at least an outgoing link in the backbone) are realizations of independent beta-distributed random variables with mean 0.71 and variance 0.01. Finally, the activity potentials are selected as realizations of independent and identically power-law distributed random variables with exponent equal to -2.5 and lower cut-off $a_{\min} = 0.01$. The resulting routing matrix P is represented through a color-coded graph in Fig. 5b.

The system is then simulated for a time-window of duration T and the data corresponding to the integrated network are stored in the weighted adjacency matrix W.

Initially, we test the capability of our technique to correctly identify the model parameters when a sufficiently large amount of data is available. Then, we study the performance of our method in a critical scenario in which the system is observed for a short time-window and the data vector of temporal links is much smaller, in order to appreciate the effectiveness of the approach even in this situation.

In our first analysis, we set $T = 10^6$ and we observe approximately 1,500,000 temporal connections during the time-window. We apply our parameter identification technique by setting the community belief parameter γ_i equal

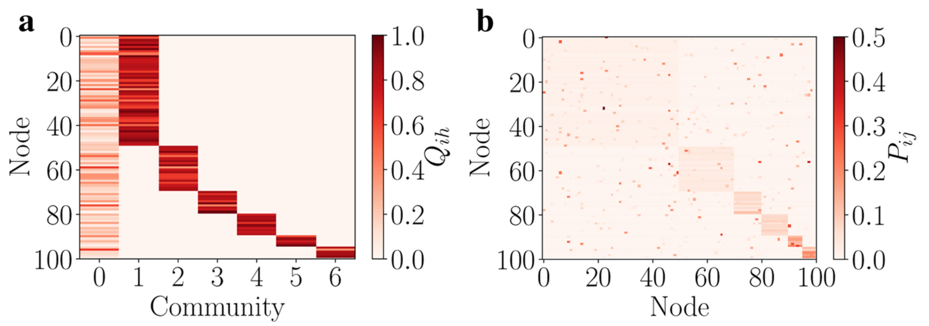

Fig. 5 Matrices characterizing the benchmark with exclusive communities. In (**a**), the community strength matrix Q is represented, while (**b**) shows the resulting routing matrix P. The values of the entries of the matrices are represented by the corresponding color intensity

to the maximum value of the confidence interval identified through the procedure proposed in "Confidence interval for the community belief parameter" section, i.e., $\gamma_i = \bar{\gamma}_i, \forall i \in \mathcal{V}$. This choice is the one that allows us to explain the largest part of the link generation process in terms of communities, compatibly with the observed data. Figure 6 illustrates the accuracy of our method in the estimation of matrix Q. We observe that the nonzero entries of matrix Q are estimated with a high accuracy, and without any bias related to the size of the communities and the heterogeneity in the parameters. In fact, accuracy does not change over the different communities, associated with different colors in Fig. 6, and it is high both for small and for large values of Q_{ij} and R_{ij}. Specifically, the mean square error of the estimated entries of matrix Q varies from 0.011 to 0.020 over the communities, with no statistically significant difference between them, while for the entries of matrix R, it is equal to 0.012. In additional simulations, here omitted for brevity, we have observed that increasing the number of communities k has no significant effect on the performance of our parameter identification method.

In our second analysis, we reduce the size of the data vector, setting $T = 10^5$ and $T = 10^4$, and generating approximately 150,000 and 15,000 temporal connections, respectively. Figure 7 reveals that, when the system can be observed only for a short time-window and few temporal links are observed, the choice of the largest value of the community belief γ_i in its confidence interval may lead to an overestimation of the contribution of the communities in the link generation process. We observe from Fig. 7a-b that the entries of vector λ, which weight the contribution of communities in the link generation mechanism, tend to be overestimated for small values of T. In fact, the mean square error of the estimated entries of vector λ increases from 0.020 for $T = 10^6$, to 0.055 and 0.185, for $T = 10^5$ and $T = 10^4$, respectively. We believe that the overestimation of the community weights is a common phenomenon when the number of temporal links is small. To address this issue, we suggest to reduce the value of the parameter γ_i, within the confidence interval established in "Confidence interval for the community belief parameter" section. Figure 8 shows that, also with few data available, an accurate estimation of the vector λ and a good estimation of the community strength matrix Q can be obtained by selecting smaller values for the community belief parameter vector. In Fig. 8a we can appreciate an excellent agreement between the estimated community weight vector and

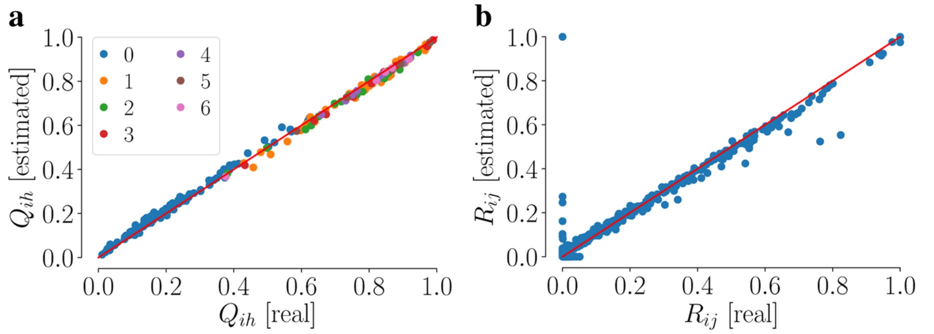

Fig. 6 Validation on the synthetic benchmark with exclusive communities. Estimation of the nonzero entries of (**a**) the community matrix Q and (**b**) the backbone matrix R for the benchmark with exclusive communities, from data sampled over a time-window of duration $T = 10^6$. The parameters are set as $\gamma_i = \bar{\gamma}_i, \forall i \in \mathcal{V}$. Different colors in (**a**) refer to the different communities

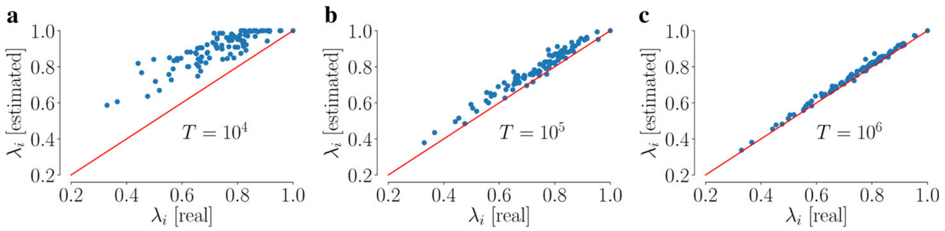

Fig. 7 Dependence of the identification performance on the duration of the time-window T. Estimation of the entries of the weights vector λ for increasing duration of the time-window T, with $\gamma_i = \bar{\gamma}_i$, $\forall i \in \mathcal{V}$. For small values of T, setting $\gamma_i = \bar{\gamma}_i$ seems to yield an overestimation of the contribution of communities in the link generation mechanism

the actual one (mean square error equal to 0.003). In Fig. 8b we observe that, even though the accuracy in the estimation of the matrix Q is reduced with respect to the case with large T, there is still a satisfactory agreement between the estimated entries of the community strength matrix and the corresponding real quantities (mean square error equal to 0.091). In our simulations, we select the value for the community belief parameter vector γ by performing a bisection method in the range $\gamma_i \in [0, \bar{\gamma}_i]$, to minimize the absolute deviation between the estimated vector λ and the original benchmark community weight vector. This confirms our intuition that the best model estimation is within the confidence interval we have assessed. In our example, we observe that, when the temporal link set is reduced to the 1% of the original amount (i.e., 15,000 links), the optimal value for gamma is found to be in average the 18% of the extreme value $\bar{\gamma}$. In real-world scenarios, where the real values of λ are unknown, the optimal selection of parameter γ remains an open problem, which will be tackled in our future research.

Hierarchical communities

Here, we assess our method on a second benchmark in which communities present a hierarchical structure. Specifically, we define a two-level hierarchy: nodes are first partitioned into two first-level communities, then, each of these communities is split into two second-level communities. Each node thus belongs to a first-level community and to a second-level one, besides the all-to-all community. Details are reported in Table 3.

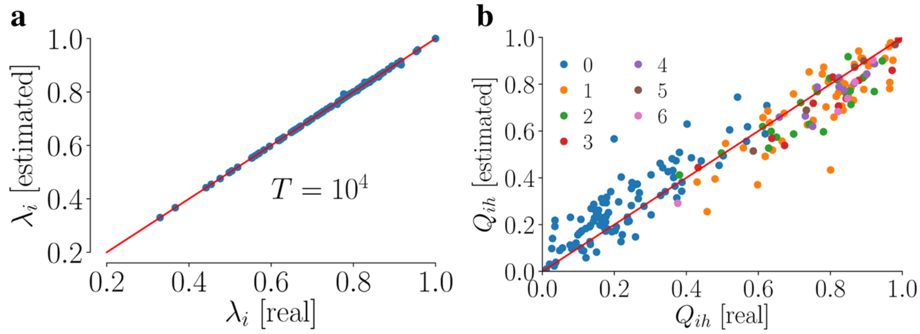

Fig. 8 Identification performance for T small, reducing γ_i. Estimation of (**a**) the vector λ and (**b**) the matrix Q for $T = 10^4$, using the optimal value of $\gamma_i \in [0, \bar{\gamma}_i]$, $i \in \mathcal{V}$, selected using a bisection method. The optimal value of γ_i is in average 18% of the extreme value $\bar{\gamma}_i$

Table 3 Benchmark with hierarchical community structures

# community	Size n_h	Members $\mathcal{V}_h$
1	50	$\{1,\dots,50\}$
2	50	$\{51,\dots,100\}$
3	25	$\{1,\dots,25\}$
4	25	$\{26,\dots,50\}$
5	25	$\{51,\dots,75\}$
6	25	$\{76,\dots,100\}$

Matrix Q is generated similarly to the previous benchmark. The entries of its first column are selected from a beta distribution with mean 0.25 and variance 0.02, each one independent of the others. The other two nonzero terms of each row are realizations of uniformly distributed random variables, normalized to obtain a stochastic matrix. Matrix Q obtained according to this procedure is illustrated, through color coding, in Fig. 9a. Then, the backbone matrix R, the weight vector λ, and the activity rate vector a are generated following the same procedure described in the previous section.

The system is simulated for a time-window of duration $T = 10^6$, obtaining approximately 1,500,000 temporal connections, the weighted adjacency matrix W of the integrated network is generated, and our technique is used to estimate the parameters. The results of our analysis, illustrated in Fig. 7b-c, suggest that our method is also able to deal with hierarchical community structures. We observe that we are able to identify the model parameters with a high accuracy and without any bias due to the different levels in the hierarchical community structure. In fact, in Fig. 9c, we observe that there is no significant difference in the accuracy of the estimation of the entries corresponding to the first-level communities (i.e., 1 and 2) and the second-level ones (i.e., 3–6). Specifically, the mean square error of the estimated entries of matrix Q varies over the communities from 0.015 to 0.025 (in average, it is equal to 0.021 for the first-level communities and 0.017 for the second-level ones, with no statistically significant difference between the two quantities). Finally, we observe that, also in this case where hierarchical communities are present, the problem of a reduced size of the data vector can be addressed by reducing the tradeoff parameter vector γ, in order to avoid data overfitting. Results are omitted for brevity.

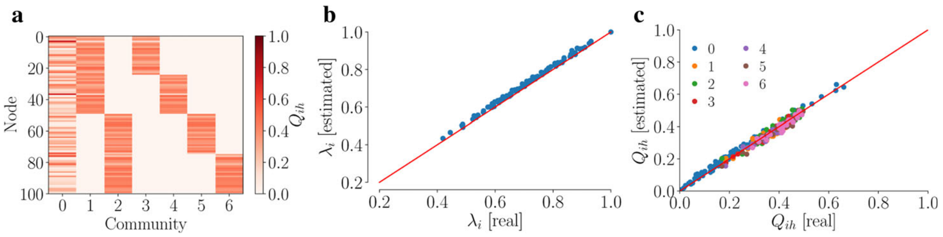

Fig. 9 Validation on a synthetic network with hierarchical communities. In (**a**), matrix Q is illustrated (color intensity is proportional to the value of the entry). Estimations of the vector λ and of the matrix Q are shown in (**b**) and (**c**), respectively. Different colors in (**c**) indicate different communities

Overlapping communities

Finally, we consider a benchmark in which communities have an overlapping structure. The $n = 100$ nodes are divided into 7 communities, as detailed in Table 4.

The community structure presents several overlaps, notably between communities 2,3,4, and 7. The community strength matrix Q (illustrated in Fig. 10a), the backbone matrix R, as well as the two vectors λ and a are defined following the procedure presented in the previous benchmarks. Then, the system is simulated for a time-window of duration $T = 10^6$, generating approximately 1,500,000 temporal connections, and our technique is used to identify the parameters from the weighted adjacency matrix W of the integrated network obtained from our simulations. Also in this scenario, as illustrated in Fig. 10b and c, our method is able to identify the model parameters with high accuracy and is free of any bias due to the presence of overlaps between the communities. In fact, in Fig. 10b we observe that the accuracy in the estimation of the entries of the community weight vector λ are not influenced on whether a node belongs to an overlapping community or not: the average mean square error for the nodes in overlapping communities is equal to 0.019, while for nodes not in overlapping communities is equal to 0.015, with no statistically significant difference between the two quantities. Similarly, also the outcome of the estimation of the community strength matrix Q is not influenced by the presence of overlaps between the communities, as can be observed in Fig. 10c, by comparing nodes that belong to different communities. In fact, the mean square error of the estimated entries of matrix Q varies over the communities from 0.013 to 0.023 (in average, it is equal to 0.015 for the communities with no overlaps and 0.017 for the overlapping ones, with no statistically significant difference between the two quantities). Also in this case, the tradeoff parameter vector γ can be set to a smaller value than the extreme $\bar{\gamma}$ to avoid overestimation of λ when little data is available, similarly to what discussed in the case of exclusive communities.

Case studies

Enron email corpus

We use our method to enrich the results of community detection for a real-world case study: the Enron email corpus (Cohen). This is a dataset of more than 500,000 emails sent by the 158 employees of Enron company from 1979-12-31 to 2002-06-21, when the company failed. In order to deal with a uniform dataset, in which the community structure and the irreducible backbone can be assumed to be constant, we restrict the dataset to the portion of mails sent after 1998-11-13. We also remove self-sent emails and nodes that do not send or receive any email. After such a data cleaning procedure, we obtain a dataset

Table 4 Benchmark with overlapping community structures

# community	Size n_h	Members $\mathcal{V}_h$
1	50	$\{1,\ldots,50\}$
2	25	$\{46,\ldots,70\}$
3	15	$\{71,\ldots,85\}$
4	10	$\{81,\ldots,90\}$
5	5	$\{91,\ldots,95\}$
6	5	$\{96,\ldots,100\}$
7	45	$\{41,\ldots,85\}$

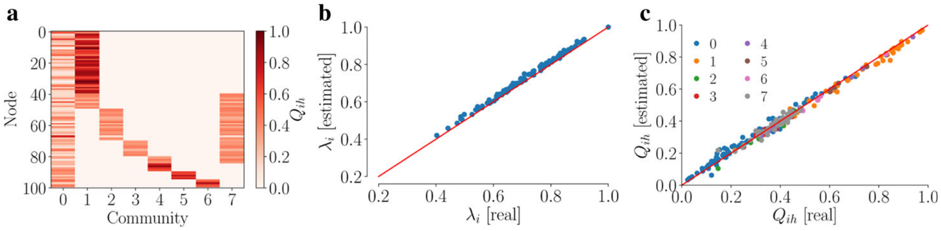

Fig. 10 Validation on a synthetic network with overlapping communities. In (**a**), matrix Q is illustrated (color intensity is proportional to the value of the entry). Estimations of the vector λ and of the matrix Q are shown in (**b**) and (**c**), respectively. Different colors in (**c**) indicate different communities

with $n = 143$ employees and 108,786 emails, which identify a temporal network where each employee is a node and each email determines a link from the sender to the receiver.

We then use a community detection algorithm on the integrated network to identify the community structure. We observe that the application of different algorithms for community detection may lead to the identification of different community structures. As stated in the introduction, our method can be used to establish a criterion to discriminate among the outcome of different community detection algorithms. In fact, for any community structure obtained by means of a different community detection algorithm, we can identify the model parameters using our method, and then compare the average entry of the community weight vector λ over the nodes, namely

$$< \lambda >:= \frac{1}{n} \sum_{i \in \mathcal{V}} \lambda_i. \tag{21}$$

This quantity measures the fraction of links that can be statistically explained by means of a community-based mechanism. Therefore, the community structure that is able to produce the highest value of $< \lambda >$ is the one that is able to explain the largest part of the link formation process.

In our case study, we apply three different algorithms to detect the communities from the integrated network: Infomap (Rosvall and Bergstrom 2008), Louvain (Blondel et al. 2008), and OSLOM (Lancichinetti et al. 2011). Since the community detection algorithms are based on randomized techniques, we perform 100 runs of each algorithm. For each of these 300 outputs, we perform our parameter identification method. The community belief vector γ is chosen within its confidence interval by using a bisection method to maximize the value of $< \lambda >$. Then, for each of the three community detection algorithms, we select the run that yields the largest value of $< \lambda >$. The estimated matrices Q for the three different algorithms are illustrated and compared in Fig. 11. We observe that the three outputs are significantly dissimilar, since a different number of communities is originally detected. Specifically, we observe that the largest community identified by Louvain is split into two or more small communities by the other algorithms. Despite these differences, we can identify some common patterns. For instance, there is a first group of 17 nodes that belong to a "strong" community, where members have high tendency of generating inter-community links. This feature of the system emerges from all of the three outputs, as can be observed by comparing the different panels of Fig. 11.

When comparing the output of the different community detection algorithms used in this case study, we observe that, interestingly, Infomap produces a very stable result:

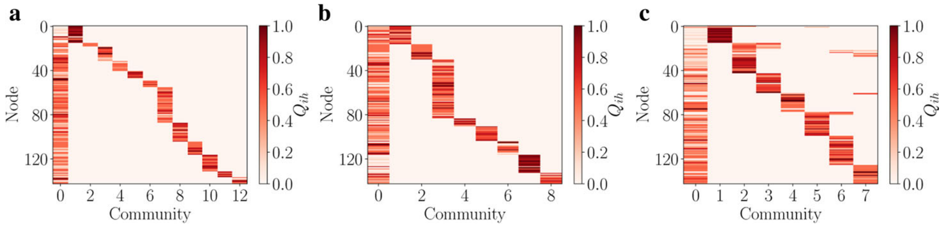

Fig. 11 Results of our parameter identification method on the Enron email corpus case study. Matrix *Q* is estimated, starting from the community structure detected by using (**a**) Infomap, (**b**) Louvain, and (**c**) OSLOM

in fact, in each of the runs it always retrieves the same community partition. Louvain, instead, identifies 6 different outcomes in the 100 runs, whereas OSLOM produces a different outcome in each run. In Fig. 12a, we plot the overlapping normalized mutual information (ONMI) evaluated between each pair of partitions produced by the same method (McDaid et al. 2011). This figure supports our claim that the output of OSLOM is strongly unstable, since each run produces a different outcome. In fact, the correlation between two different outputs can be small, as seen in the box-plot. Instead, the six different outputs of Louvain algorithm are strongly correlated. In Fig. 12b, the distribution of the value of $< \lambda >$ in different runs of the algorithm is illustrated. We observe that Infomap outperforms the other two algorithms, while OSLOM and Louvain seem to have a similar performance. Finally, in Fig. 12c we plot the ONMI of an OSLOM partition with the Infomap partition as a function of the quantity $< \lambda >$. From this figure, we can observe a significant positive correlation between the two quantities, which supports our intuition that $< \lambda >$ might be used as a performance index of the community detection algorithm.

Primary school

We apply our algorithm to a second real-world case study: the SocioPattern primary school dataset (SocioPatterns). This dataset consists of a temporal network of face-to-face interactions between students and teachers in a French primary school, recorded via proximity sensors. The dataset comprises 77,602 interactions (sampled with a time resolution of 20 s) between $n = 242$ individuals over a time-window of duration $T = 2$ days.

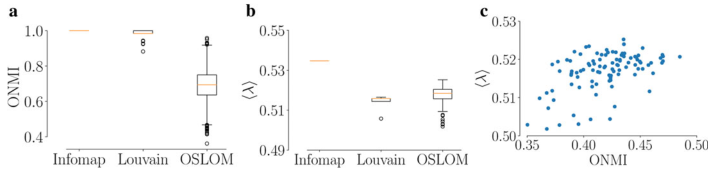

Fig. 12 Comparison between different community detection algorithms for the Enron email corpus case study. In (**a**), we show the box-plots of the distribution of the ONMI for each pair of outputs over 100 runs of the three community detection algorithms. The output of OSLOM seem to be highly unstable, confirming our preliminary observations. In (**b**), we plot the distribution of $\langle\lambda\rangle$ for the three algorithms. Infomap has the largest value of such a quantity. In (**c**), for each one of the 100 different outcomes of the OSLOM algorithm, we compare $\langle\lambda\rangle$ with the ONMI with the (unique) output of Infomap. The positive correlation that seems to be present supports our suggestion of using $\langle\lambda\rangle$ to evaluate the outcome of a community detection algorithm. In fact, the outputs that are closer to the stable community structure identified by Infomap have a higher value of $\langle\lambda\rangle$

In this dataset, individuals are naturally partitioned: 232 of them are students, divided into 10 classes, and 10 are teachers (Stehlé et al. 2011; Gemmetto et al. 2014). These metadata provide a ground truth for the community structure. The main limitation of this dataset is that the direction of the interactions is not known, since it cannot be registered by the proximity sensors. In order to apply our method, in the absence of exact information on the link direction, we assume it to be homogeneously distributed. Hence, we perform a Monte Carlo parameter identification over 100 runs in which we randomize over the direction of each link in the dataset. Specifically, for each undirected link $\{i, j\}$, we interpret it as a directed link from i to j with probability $1/2$, and as a directed link from j to i, otherwise, each one independent of the others. Since classes provide an evidence on the ground truth of the community structure and the number of interactions in the dataset is sufficiently large, we set the largest value of belief parameter within its confidence interval, i.e., $\gamma_i = \bar{\gamma}_i$, for all $i \in \mathcal{V}$. Then, we evaluate the average vector $\hat{\lambda}$ and the average matrices $\hat{Q}$ and $\hat{R}$ over the multiple realizations.

The results illustrated in Fig. 13 show that our model is able to capture the community structure of the system, supporting the hypothesis that comes from the natural partition of students into their classes. In fact, the distribution of $\hat{\lambda}$ illustrated in Fig. 13a shows that the class-based community structure is able to describe a large part of the observed links for most of the nodes. This can also be observed by the large involvement of members in their communities, illustrated in Fig. 13b. Notably, the teachers, corresponding to the last row of Fig. 13b, make an exception: they have small values of $\hat{Q}$ within their community (last column), while they have a large involvement in the all-to-all community. This seems to reflect the reality, since a teacher often interacts more with students (of several classes) than with other teachers. It is worth noticing that this is an information that a traditional community detection algorithm can hardly reveal. In addition, a more detailed structure of dyadic relationships, both within and outside the classes, is revealed in the backbone matrix $\hat{R}$ represented in Fig. 13c. From its structure, one can infer the presence of strong relationships between students, mostly classmates. From these interactions one can infer the presence of subcommunities within each class and use this information to reconstruct a hierarchical community structure. It also worth noticing that, for last-years students (the last rows before teachers), the dyadic relationships in the backbone are not limited to

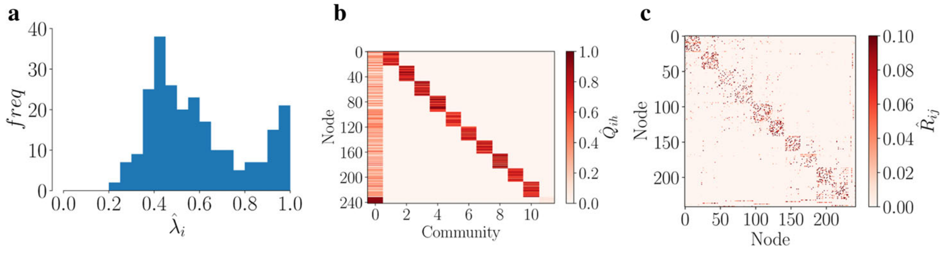

Fig. 13 Monte Carlo parameter identification (over 100 random link orientations) for the primary school case study. In (**a**), we show the distribution of the estimated parameter $\hat{\lambda}$, which quantifies the role of communities in the network formation process. In average, half of the links are generated following the community structure. In (**b**), we plot the estimated community strength matrix $\hat{Q}$. From the output, we observe that all the students have a large involvement in their classes, whereas teachers are more active in the all-to-all community than in the teacher community. In (**c**), we plot the estimated irreducible backbone of the network $\hat{R}$. We observe that the network is sparse, and most of the nonzero terms are within the classes. These dyadic relationships may represent, e.g., students who sit next to each others

the classmates, but also inter-classes nonzero entries are present. This is consistent with other analyses performed on the same dataset, which show that last-years students are more active in generating out-of-class relationships (Stehlé et al. 2011; Gemmetto et al. 2014).

Conclusion

In this paper, we deal with the problem of modeling and characterizing the complex network structure of real-world systems. First, we present a mathematical model for temporal networks that generalizes the ADN paradigm, by including link directionality, the presence of a heterogeneous, hierarchical, and overlapping community structure, and the existence of an irreducible backbone of connections. Then, based on this model, we propose a technique to estimate the model parameters from empirical data and assess the effect of communities and the irreducible backbone on the link generation process into an intelligible form, providing a mesoscopic description of the system at the communities level. The proposed technique is based on the introduction of a free parameter that can be calibrated within a confidence interval. This parameter models our belief in the role of communities in the link formation mechanism. We validate our method on three different synthetic networks and on a real-world case study, with satisfactory results. We also apply our method to two different real-world case studies. In the first one, the ground truth about the community structure is unknown and our method is used to establish a criterion to assess the performance of different community detection algorithms. In the second scenario, a ground truth about the community structure is instead provided by the partition of students and teachers in classes. In this case, we are able to *i*) retrieve the actual partition in classes and *ii*) reveal the different role of students and teachers in their classes.

The presented method is characterized by a reasonable computational effort. This property, together with the possibility of analytical treatment exhibited by ADNs, is essential to tackle real-world problems. For example, the possibility to detect the role of nodes and communities in interactions between individuals allows for the design of accurate targeted immunization strategies for the case of disease spreading (Masuda 2009; Salathè and Jones 2010; Gong et al. 2013), or for the detection of closed communities that drive the spread of misinformation and fake news, named echo chambers (Del Vicario et al. 2016). For these reasons, we believe that the possibility of unveiling the architecture of a complex system, through the characterization of its community structure, may play a fundamental role in the development of effective techniques to address real-world problems, with potential invaluable benefits to the society.

Abbreviations

ADN: activity driven network; LR: likelihood ratio; ONMI: overlapping normalized mutual information; OSLOM: order statistics local optimization method; rADN: routed activity driven network

Acknowledgment

The authors acknowledge the SocioPatterns collaboration.

Funding

This publication is realized with the contribution of the Italian Ministry of Foreign Affairs and International Cooperation, within the project "Mac2Mic", "Macro to Micro: uncovering the hidden mechanisms driving network dynamics," funded under the bilateral agreement between Italy and Israel for scientific, technological, and industrial cooperation. It is also supported by Compagnia di San Paolo and by the Italian Ministry of Education, Universities and Research under grant Dipartimenti di Eccellenza 2018–2022 (E11G18000350001).

Availability of data and materials
The datasets generated during the current study are available upon request. The source code used to perform the parameter identification is avaialble at https://github.com/cbongiorno/RADNFIT. The datasets analyzed in the case study is available in the Enron Email Dataset (https://www.cs.cmu.edu/~./enron/) and in the SocioPatterns Primary School Temporal Network Data (http://www.sociopatterns.org/datasets/primary-school-temporal-network-data/), respectively.

Authors' contributions
AR coordinated the research; CB performed the numerical studies and conceived the parameter identification method; LZ designed the model and wrote the initial draft; all authors contributed to data analysis and writing the final manuscript. All authors read and approved the final manuscript.

Competing interests
The authors declare that they have no competing interests.

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

Author details
[1]Department of Electronics and Telecommunications, Politecnico di Torino, Corso Duca degli Abruzzi, 24, 10129 Torino, Italy. [2]Department of Mathematical Sciences "G.L. Lagrange", Politecnico di Torino, Corso Duca degli Abruzzi, 24, 10129 Torino, Italy. [3]Office of Innovation, New York University Tandon School of Engineering, 11201 Brooklyn NY, US.

Received: 6 December 2018 Accepted: 6 March 2019
Published online: 29 March 2019

References
Alessandretti L, Sun K, Baronchelli A, Perra N (2017) Random walks on activity-driven networks with attractiveness. Phys Rev E 95(5):052318

Bailey NTJ (1990) The Elements of Stochastic Processes with Applications to the Natural Sciences. Wiley, New York

Ballerini M, Cabibbo N, Candelier R, Cavagna A, Cisbani E, Giardina I, Lecomte V, Orlandi A, Parisi G, Procaccini A, Viale M, Zdravkovic V (2008) Interaction ruling animal collective behavior depends on topological rather than metric distance: Evidence from a field study. Proc Natl Acad Sci USA 105(4):1232–1237

Bao W, Michailidis G (2018) Core community structure recovery and phase transition detection in temporally evolving networks. Sci Rep 8(1):12938

Benson AR, Gleich DF, Leskovec J (2016) Higher-order organization of complex networks. Science 353(6295):163–166

Bishop C (2006) Pattern Recognition and Machine Learning. Springer, New York

Blondel VD, Guillaume J-L, Lambiotte R, Lefebvre E (2008) Fast unfolding of communities in large networks. J Stat Mech Theory Exp 2008(10):10008

Bongiorno C, Zino L, Rizzo A (2018) On unveiling the community structure of temporal networks. In: Proceedings of the 57th IEEE Conference on Decision and Control (CDC). pp 6210–6215

Boyd S, Vandenberghe L (2004) Convex Optimization. Cambridge University Press, New York

Casella G, Berger RL (2002) Statistical Inference, vol. 2. Duxbury, Pacific Grove

Cohen WW Enron Email Dataset. https://www.cs.cmu.edu/~./enron/. Accessed 27 Feb 2019

Del Vicario M, Bessi A, Zollo F, Petroni F, Scala A, Caldarelli G, Stanley HE, Quattrociocchi W (2016) The spreading of misinformation online. Proc Natl Acad Sci USA 113(3):554–559

Erdős P, Rényi A (1959) On random graphs. Publ Math Debrecen 6:290–297

Estrada E (2011) The Structure of Complex Networks: Theory and Applications. Oxford University Press, Oxford

Fortunato S, Hric D (2016) Community detection in networks: A user guide. Phys Rep 659:1–44. Community detection in networks: A user guide

Gemmetto V, Cardillo A, Garlaschelli D (2017) Irreducible network backbones: unbiased graph filtering via maximum entropy. arXiv preprint arXiv:1706.00230

Gemmetto V, Barrat A, Cattuto C (2014) Mitigation of infectious disease at school: targeted class closure vs school closure,. BMC Infect Dis 14(1):695

Girvan M, Newman MEJ (2002) Community structure in social and biological networks. Proc Natl Acad Sci USA 99(12):7821–7826

Gong K, Tang M, Hui PM, Zhang HF, Younghae D, Lai Y-C (2013) An efficient immunization strategy for community networks. PLoS ONE 8(12):1–11

Hammings R (1973) Numerical Methods for Scientists and Engineers, 2nd edition. Dover Publications, New York

Holme P, Saramäki J (2012) Temporal networks. Phys Rep 519(3):97–125

Karsai M, Perra N, Vespignani A (2014) Time varying networks and the weakness of strong ties. Sci Rep 4:4001

Khan BS, Niazi MA (2017) Network community detection: A review and visual survey. arXiv preprint arXiv:1708.00977

Lancichinetti A, Fortunato S, Kertész J (2009) Detecting the overlapping and hierarchical community structure in complex networks. New J Phys 11(3):033015

Lancichinetti A, Radicchi F, Ramasco JJ, Fortunato S (2011) Finding statistically significant communities in networks. PLoS ONE 6(4):1–18

Latapy M, Viard T, Magnien C (2018) Stream graphs and link streams for the modeling of interactions over time. Soc Netw Anal Min 8(1):61

Lei Y, Jiang X, Guo Q, Ma Y, Li M, Zheng Z (2016) Contagion processes on the static and activity-driven coupling networks. Phys Rev E 93(3):032308

Leicht EA, Newman MEJ (2008) Community structure in directed networks. Phys Rev Lett 100(11):118703

Li D, Han D, Ma J, Sun M, Tian L, Khouw T, Stanley HE (2017) Opinion dynamics in activity-driven networks. EPL 120(2):28002
Liu S, Perra N, Karsai M, Vespignani A (2014) Controlling contagion processes in activity driven networks. Phys Rev Lett 112(11):118702
Masuda N (2009) Immunization of networks with community structure. New J Phys 11(12):123018
McDaid AF, Greene D, Hurley N (2011) Normalized mutual information to evaluate overlapping community finding algorithms. arXiv preprint arXiv:1110.2515
Nadini M, Rizzo A, Porfiri M (2018a) Epidemic spreading in temporal and adaptive networks with static backbone. IEEE Trans Netw Sci Eng. https://doi.org/10.1109/TNSE.2018.2885483
Nadini M, Sun K, Ubaldi E, Starnini M, Rizzo A, Perra N (2018b) Epidemic spreading in modular time-varying networks. Sci Rep 8(1):2352
Newman MEJ (2003) The structure and function of complex networks. SIAM Rev 45(2):167–256
Newman MEJ (2006) Modularity and community structure in networks. Proc Natl Acad Sci USA 103(23):8577–8582
Onnela J-P, Saramäki J, Hyvönen J, Szabó G, Lazer D, Kaski K, Kertész J, Barabási A-L (2007) Structure and tie strengths in mobile communication networks. Proc Natl Acad Sci USA 104(18):7332–7336
Palla G, Barabási A-L, Vicsek T (2007) Quantifying social group evolution. Nature 446:664–667
Palla G, Derényi I, Farkas I, Vicsek T (2005) Uncovering the overlapping community structure of complex networks in nature and society. Nature 435:814–818
Pastor-Satorras R, Castellano C, Van Mieghem P, Vespignani A (2015) Epidemic processes in complex networks. Rev Mod Phys 87(3):925
Perra N, Gonçalves B, Pastor-Satorras R, Vespignani A (2012) Activity driven modeling of time varying networks. Sci Rep 2:469
Petri G, Barrat A (2018) Simplicial activity driven model. Phys Rev Lett 121(22):228301
Pons P, Latapy M (2011) Post-processing hierarchical community structures: Quality improvements and multi-scale view. Theor Comput Sci 412(8):892–900
Pozzana I, Sun K, Perra N (2017) Epidemic spreading on activity-driven networks with attractiveness. Phys Rev E 96(4):042310
Ravasz E, Somera AL, Mongru DA, Oltvai ZN, Barabási A-L (2002) Hierarchical organization of modularity in metabolic networks. Science 297(5586):1551–1555
Ribeiro B, Perra N, Baronchelli A (2013) Quantifying the effect of temporal resolution on time-varying networks,. Sci Rep 3:3006
Rizzo A, Porfiri M (2016) Innovation diffusion on time-varying activity driven networks. Eur Phys J B 89:20
Rizzo A, Frasca M, Porfiri M (2014) Effect of individual behavior on epidemic spreading in activity driven networks. Phys Rev E 90(4):042801
Rizzo A, Pedalino B, Porfiri M (2016) A network model for ebola spreading. J Theor Biol 394:212–222
Ross SM (2009) Introduction to Probability Models. Academic Press, Cambridge
Rossetti G, Cazabet R (2018) Community discovery in dynamic networks: a survey. ACM Comput Surv 51(2):35
Rosvall M, Bergstrom CT (2008) Maps of random walks on complex networks reveal community structure. Proc Natl Acad Sci USA 105(4):1118–1123
Salathè M, Jones JH (2010) Dynamics and control of diseases in networks with community structure. PLOS Comput Biol 6(4):1–11
Schaub MT, Delvenne J-C, Rosvall M, Lambiotte R (2017) The many facets of community detection in complex networks. Appl Netw Sci 2(1):4
SocioPatterns SocioPatterns Primary School Temporal Network Data. http://www.sociopatterns.org/datasets/primary-school-temporal-network-data/. Accessed 27 Feb 2019
Starnini M, Pastor-Satorras R (2014) Temporal percolation in activity-driven networks. Phys Rev E 89(3):032807
Stehlé J, Voirin N, Barrat A, Cattuto C, Isella L, Pinton J, Quaggiotto M, Van den Broeck W, Régis C, Lina B, Vanhems P (2011) High-resolution measurements of face-to-face contact patterns in a primary school. PLoS ONE 6(8):23176
Sun K, Baronchelli A, Perra N (2015) Contrasting effects of strong ties on sir and sis processes in temporal networks. Eur Phys J B 88:326
Volz E, Meyers LA (2008) Epidemic thresholds in dynamic contact networks. J Royal Soc Interface 6(32):233–241
Yang J, Leskovec J (2014) Overlapping communities explain core–periphery organization of networks. Proc IEEE 102(12):1892–1902
Zhang X, Ma Z, Zhang Z, Sun Q, Yan J (2018) A review of community detection algorithms based on modularity optimization. J Phys Conf Ser 1069(1):012123
Zino L, Rizzo A, Porfiri M (2016) Continuous-time discrete-distribution theory for activity-driven networks. Phys Rev Lett 117(22):228302
Zino L, Rizzo A, Porfiri M (2017) An analytical framework for the study of epidemic models on activity driven networks. J Complex Netw 5(6):924–952
Zino L, Rizzo A, Porfiri M (2018) Modeling memory effects in activity-driven networks. SIAM J Appl Dyn Syst 17(4):2830–2854