Generating surrogate temporal networks from mesoscale building blocks

Check for updates

Giulia Cencetti ✉ & Alain Barrat

Surrogate networks can constitute suitable replacements for real networks, in particular to study dynamical processes on networks, when only incomplete or limited datasets are available. As empirical datasets most often present complex features and interplays between structure and temporal evolution, creating surrogate data is however a challenging task, in particular for data describing time-resolved interactions between agents. Here we propose a method to generate surrogate temporal networks that mimic such observed datasets. The method is based on a decomposition of original datasets into temporal subnetworks encoding local structures on a short time scale. These are used as building blocks to generate new synthetic temporal networks that will hence inherit the shape of local interactions from the datasets. Moreover, we take into account larger scale correlations on structural and temporal dimension, using them to inform the process of assembling the building blocks. We showcase the method by generating surrogate networks for several datasets of social interactions and comparing them to the original data. First, we show that surrogate data possess complex structural and temporal features similar to the ones of the original data. Second, we simulate several dynamical processes and compare their outcome on the generated and original datasets.

Networks are a reference representation tool in the field of complex systems, well suited to describe systems composed of multiple interacting agents[1–4]. Formed by a set of discrete nodes and the connections between them, networks (or graphs) schematize the existing interactions among elements, providing a representative picture of the system architecture in many disciplines, from physics to sociology, biology, and economy. In many cases, agents' interactions undergo a temporal evolution, with links appearing and disappearing over time, and their description necessitates the use of temporal networks representations and tools[5,6]. This framework proves valuable in many settings, like neuronal functions and ecosystems, but is particularly useful to describe social contexts, where connections among people change over time, both in physical and remote interactions, with non-trivial temporal and structural correlations. Both static and temporal network representations allow moreover to describe the unfolding of dynamical processes among agents such as, for instance, spreading phenomena (of diseases, opinions or information), transportation models, communication, synchronization, and consensus formation[1]. The collective behaviors resulting from these processes depend on both the structure of the underlying network and its temporal evolution[1,5,7,8], so that the correct description of their properties requires their study and simulation on temporal networks with realistic structural and temporal patterns.

However, information about real temporal sequences of interactions between agents is often incomplete, limited in size and duration, due to the many challenges in collecting datasets[5,9,10]. It is thus not straightforward to study processes on empirical temporal networks of large size and of durations long enough with respect to the processes' timescales. In this context, synthetic networks that mimic the observed complex patterns of real structures can serve as surrogate substrates on which to simulate processes. In particular, the techniques to obtain synthetic networks are not necessarily limited to creating data of the same temporal length as the empirical datasets, and can thus provide solutions to the problem of data with limited duration. For instance, it is possible to generate many instances of surrogate datasets (of the same duration as the original data) and arrange them in temporal succession to create a dataset with arbitrary duration[11,12].

Generating synthetic temporal networks with realistic features and arbitrary sizes and timescales remains however a challenging task: the interplay between temporal evolution (e.g., how each connection changes in time, the durations of interactions, the frequency at which new ones appear and existing ones disappear) and topological organization (the structural correlations between the instantaneous relations among nodes) leads indeed to convoluted spatio-temporal structures that require specifically designed tools to analyze and are difficult to reproduce[13–19].

Aix Marseille Univ, Université de Toulon, CNRS, CPT, Marseille, France. ✉e-mail: giuliacencetti@gmail.com

The methods for temporal network generation can be broadly divided into two categories: the first is represented by theoretical models[7], which start from governing rules assumed to be the basis for the evolution mechanisms of the agents' interactions. These rules are thus used to build connections among nodes and determine their temporal evolution. A well-known example is constituted by the activity driven models[20–24]: in their simplest versions, they focus on reproducing the heterogeneity between nodes' activations, without reproducing complex temporal properties. Refined versions show that short and long-term memory effects in the activation of nodes and links are needed to reproduce temporal features such as the heterogeneity of contact durations and of the times between subsequent contacts, or the emergence of correlated evolution patterns of groups of nodes (communities)[18,21,24–33]. Note that these models are typically not targeted at mimicking a specific observed network: they rather aim at investigating and validating the possible underlying mechanisms of temporal network dynamics, by trying to recreate some general features considered as relevant. In this respect, their goal is usually not to reproduce the full complexity of empirical temporal networks, but to address the emergence of interesting statistical characteristics, such as, e.g., bursty distributions of intercontact durations, or broad distributions of contact durations, using simple rules. These models also remain general, in that they do not aim at producing networks similar to data collected in a specific context (such as, e.g., the temporal network of contacts between students in a school), and are thus not suitable to produce surrogate networks of given datasets, for use in numerical simulations of processes.

The second category of models is represented by emulative algorithms that are instead specifically designed to produce surrogate networks of given datasets[10,12,34–38]. In these somewhat more applicative procedures, one starts from a specific input network and tries to create a surrogate network that mimics the input structure, i.e., that produces patterns similar to the ones observed in the input (machine learning processes can also be used for this goal[35]). These models do not make any theoretical assumption about the underlying mechanisms that led to the emergence of these patterns, nor take these mechanisms into account in the creation of the surrogate data.

Emulative algorithms have been less developed and explored than theoretical models, and their use has mostly been considered and validated in the context of numerical simulations of models of infectious diseases. Here, we give a significant twofold contribution to this field: first, we put forward a novel methodology to create surrogate temporal network datasets for a variety of contexts, which brings together the emulative and theoretical points of view; second, we show the possibility to use the obtained surrogate data to simulate a variety of dynamical processes with very different properties, describing respectively epidemic spread, opinion formation and emergence of norms in a population. The method we propose can on the one hand be ascribed to the category of emulative algorithms, as it aims at producing surrogates of given empirical temporal network datasets. On the other hand, it also leverages several theoretical assumptions on the mechanisms for the establishment and dynamics of links. First, similarly to models in which the activity of nodes depends on their recent prior interactions[24–26,39], it focuses on the behavior of individual nodes, on their recent past and on their local neighborhood to determine their future interactions. Such approach was developed by Longa et al.[40] and was shown to reproduce several time averaged quantities (such as the number of instantaneous neighbors of a node, averaged over nodes and time) and instantaneous ones (such as the total number of interactions at each time). It however fails to reproduce other relevant properties observed in empirical data, such as the large values of the clustering[4] (i.e., the fraction of completely connected triads of nodes), the organization of nodes into groups[4,41], and the heterogeneity in the overall activity of connections (the total time in interaction of pairs of nodes can vary over orders of magnitude[14,42]). To overcome these limitations, we introduce the second assumption: that some network features observed in real data result from meso-scale topological correlations and long-term temporal correlations. We thus include mechanisms to reproduce global and meso-scale properties observed in the real temporal

networks that we want to mimic. Notably, the technique that we develop makes it possible to generate a network of arbitrary length, not limited by the temporal length of the original dataset it is based on.

In the following, we first provide a detailed explanation of the methodology. We use it to produce surrogate datasets of eight empirical temporal networks describing social interactions in different contexts, namely face-to-face contacts in a primary school[43], a high school[44,45], a workplace[46], and a conference[14], and co-location contacts in a primary school and a middle school[47] and a university campus[48]. The surrogates are first generated with the same temporal length as the original datasets in order to test the model. We assess the similarity between the surrogate and original datasets along a series of structural and temporal characterization tools of temporal networks. We then consider three dynamical processes with different dynamical properties, to showcase the possibility to use such surrogate data in numerical studies of these processes: a generic model for the spread of infectious diseases[1,8], the Deffuant model for opinion dynamics[49,50], and the Naming Game model for the emergence of conventions in a population[51,52]. We also show the results obtained with two simplified versions of the method, which allow us to outline the role and the importance of the different mechanisms at play in the generation of the surrogate data. We moreover go one step further and perform numerical experiments using surrogate data generated on longer timescales than the observed network's temporal length.

Results
Generating surrogate temporal networks
Let us consider a temporal network in discrete time, with N nodes and T temporal snapshots. There are in principle no constraints on the characteristics of this network, that we denote as the "original network" to distinguish it from the synthetic surrogate networks that our method will generate. We first explain the procedure to generate a network with the same number of nodes and the same temporal length as the original and we will discuss later how to adapt the methodology to generate longer networks.

Building blocks of the original network. The generation process begins with the extraction of information from the (known) original network. The first step consists in decomposing the original network into building blocks, called "Egocentric Temporal Neighborhoods", which describe how each node interacts with its neighborhood on a short time scale[53]. The Egocentric Temporal Neighborhood of a node i at time t is indeed defined as the temporal subgraph of the original network composed by i, all the nodes that interact with it between t and $t + d$, and their links with i. The connections between the neighbors of i are thus not included (see top panel of Fig. 1). The time length of the subnetwork is given by $d + 1$, where d is a free parameter of the method. In Fig. 1 and in all the successive figures we use $d = 2$, a comparison between different values of d is provided in the Supplementary Note 1.

Next, each Egocentric Temporal Neighborhood is mapped onto an *ego-subgraph*, which encodes just the shape of the local interactions: each ego-subgraph is indeed the equivalence class of all isomorphic Egocentric Temporal Neighborhoods (just as a motif in a static network is an equivalence class of isomorphic static subgraphs). To obtain the ego-subgraph, one thus removes the node identities of the Egocentric Temporal Neighborhood, but one keeps the crucial information about (i) which node is the central one (ego) and (ii) which nodes are neighbors of ego in different snapshots. The ego-subgraphs are thus not simply successions of stars in the different snapshots. For instance, in the sketch of Fig. 1a, the neighbor drawn to the top left of the ego is connected to it in the first two snapshots, and this information is kept in the resulting ego-subgraph (rightmost sketch of Fig. 1a). Note that, for computational ease of manipulation, each ego-subgraph can be mapped to a binary string, with the following procedure[53]: for each neighbor of the ego, we create a binary sequence of length $d + 1$ with a 1 each time the neighbor is connected to the ego, and a 0 when it is not (in the example of Fig. 1, there are thus 5 sequences of length 3: for instance, the sequence for the neighbor to the top left of the ego is 110, which encodes the information that it is connected to ego in the two first snapshots but not
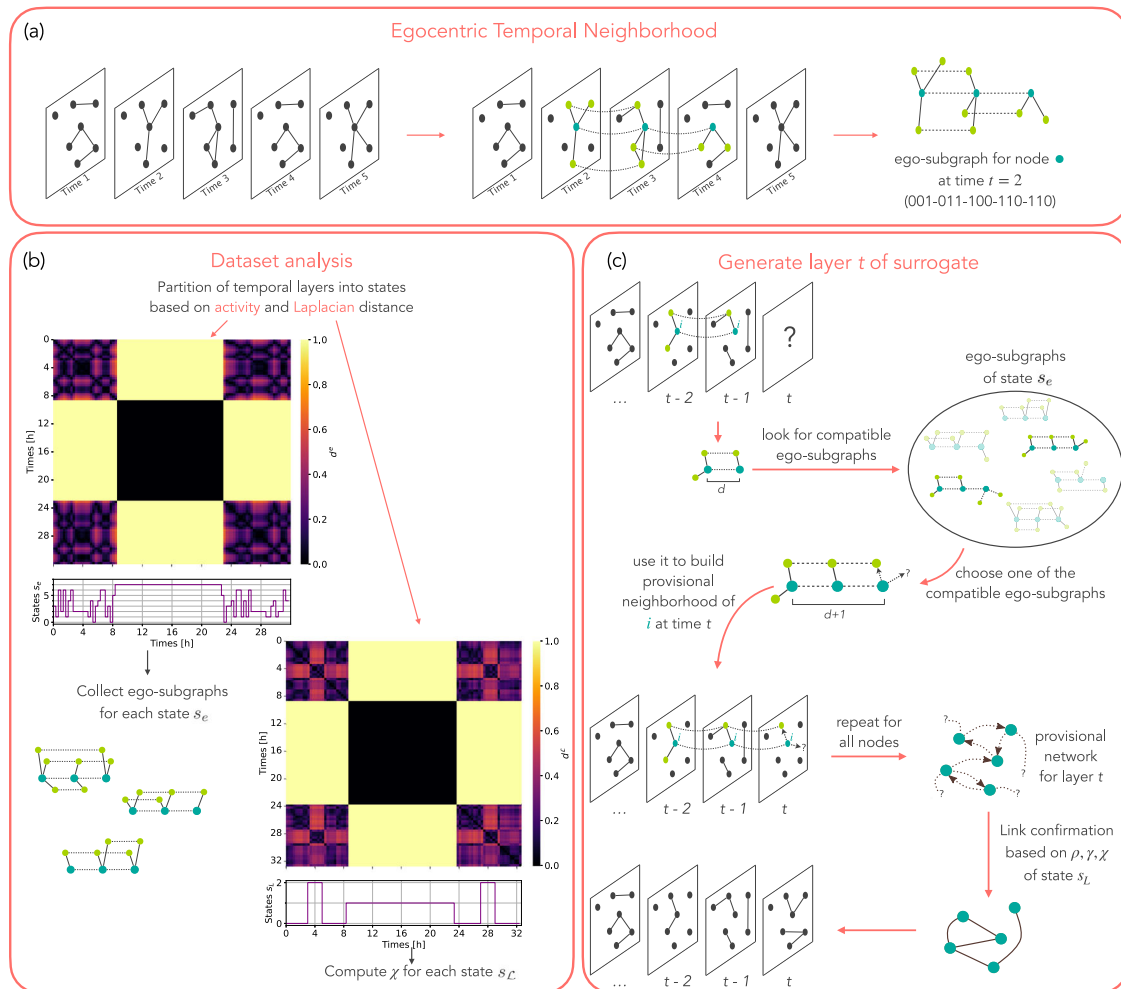
Fig. 1 | Schematic procedure for generating surrogates. a Example of Egocentric Temporal Neighborhood extraction from a temporal network. An ego-node (in aqua green) and its neighbors (in light green) are highlighted for three time steps ($d = 2$) and the corresponding ego-subgraph is depicted on the right together with its univocal signature. Note that, even if the ego-subgraph does not contain information on the identity of the nodes (as it is an equivalence class of all isomorphic Egocentric Temporal Neighborhoods), it retains the information on whether some neighbors of the central node (the ego) are the same in different snapshots. **b** The temporal snapshots are partitioned into states based on d^e (left) and $d^{\mathcal{L}}$ (right). The two matrices report the distances among snapshots for a temporal network describing interactions between children in a primary school[43], with temporal resolution of 5 min. For d^e we use here k_d with $d = 2$. The plot below each matrix

depicts the succession of states associated to each snapshot (the state numbers do not carry any specific meaning, they just serve to label the snapshots). **c** Generating a generic snapshot t of the surrogate involves several steps. For each node i (here the aqua green node), we first extract its Egocentric Temporal Neighborhood on the surrogate snapshots $t - d, \cdots, t - 1$ previously constructed, and build the resulting ego-subgraph g_i^d of temporal length d; we extract at random, from the set of ego-subgraphs of length $d + 1$ of state $s_e \ni t$ (collected from the original dataset), an ego-subgraph g_i^{d+1} of length $d + 1$ whose d first temporal steps are isomorphic to g_i^d. The $d + 1$ snapshot of g_i^{d+1} is used to design the provisional neighborhood of i. The provisional static directed network of snapshot t is obtained by repeating this procedure for all nodes. The link confirmation step finally yields the surrogate snapshot t.

in the third); we then concatenate these sequences (using the lexicographic order).

We thus extract the Egocentric Temporal Neighborhood and the resulting ego-subgraph of each node and at each time. Note that, since the node identities are removed in the ego-subgraph, different Egocentric Temporal Neighborhoods (of different nodes and at different times) can yield the same ego-subgraph. Overall, the set of all the ego-subgraphs and the numbers of Egocentric Temporal Neighborhoods to which they correspond encode the local interactions of the original network, i.e., how the nodes' neighborhoods are formed and evolve on a short time-scale[53], and it has been shown that adjusting a model's parameters to reproduce the ego-subgraphs relative frequencies allows to reproduce as well several other temporal properties of the data[24].

Decomposition in temporal states. While the global set of ego-subgraphs provides a representation of all the local interactions in the temporal network, it is not suited to describe several meso-scale

structural features of the temporal snapshots seen as static networks, such as the existence of triangles or of densely connected groups of nodes. It does not either give direct insights into the temporal variations that can characterize a temporal network at various temporal scales. It cannot, for instance, differentiate between periods of high and low activity, or between periods in which nodes form separate groups and others in which they all mix together.

To understand the network's temporal organizaton, we resort to the method of [15] to gather the temporal snapshots of the original networks into groups of snapshots with similar structure: each group of snapshots then defines a "state" of the temporal network. We first assess the similarity between each pair of snapshots (t_1, t_2) through the Laplacian distance

$$d^{\mathcal{L}}_{t_1, t_2} = \sqrt{\frac{\sum_{n=1}^{N} [\lambda_n(t_1) - \lambda_n(t_2)]^2}{\max\left[\sum_{n=1}^{N} \lambda_n(t_1)^2, \sum_{n=1}^{N} \lambda_n(t_2)^2\right]}}, \quad (1)$$

where $\lambda_n(t)$ is the nth largest eigenvalue of the Laplacian $\mathcal{L}(t)$ of snapshot t[4,15]. The matrix of distances between all pairs of snapshots of an example dataset (contacts recorded between students of an elementary school during two days[43]) are shown in the left part of Fig. 1b.

Using this matrix of distances, we then cluster the temporal snapshots into states[15]. Such optimal classification is found using the Dunn's index[54] (see the Methods section "Temporal states clustering"). For the example dataset considered in Fig. 1, three states are found (corresponding in the school to the time spans of the lectures, of the lunch breaks and of the night). As we have used the Laplacian distance to determine these states, we call them "Laplacian temporal states", $s_\mathcal{L}$. They are characterized by a different organization of nodes into communities. This can be shown by building, for each state, a weighted static network aggregating all the temporal snapshots of that state (i.e., where each link is weighted by the number of snapshots in which it appears), and examining the community structures of these static networks. In the example of the school dataset of Fig. 1, a simple application of the Louvain algorithm[55] yields a partition in 9 groups for the state 0 (lectures), which approximately reflects the students school classes[43]. The resulting modularity index is high, equal to 0.74. The partition that is obtained for the state formed by the lunch breaks snapshots consists instead in 5 groups (modularity 0.38). To summarize these differences, we compute for each state $s_\mathcal{L}$ an index χ, corresponding to the ratio of inter-group and intra-group links, (see the Methods section "Modularity and value χ" for the precise definition). For the primary school example, $\chi = 0.014$ for the state corresponding to lectures and $\chi = 0.15$ for the state of the lunch breaks ($\chi = 0$ for the night, where no links are present).

We notice however that the number of interactions of singular snapshots can undergo considerable variations inside a state $s_\mathcal{L}$. It is thus also useful to define a more fine-grained temporal state decomposition, based on nodes activity. To this aim, we define $\langle k_d(t) \rangle$ as the average number of a node's distinct neighbors between t and $t + d$ (this quantity is proportional to the mean length of all ego-subgraphs at time t), and we compute the distance

$$d_{t_1,t_2}^e = \frac{|\langle k_d(t_1) \rangle - \langle k_d(t_2) \rangle|}{\langle k_d(t_1) \rangle + \langle k_d(t_2) \rangle]} , \qquad (2)$$

between all pairs of times in $[1, T]$. The resulting matrix of distances in shown in Fig. 1 for the primary school dataset. Using this matrix to cluster the snapshots and extract states of the temporal network results here in 8 states s_e. As these states correspond to different levels of activity, and as the ego-subgraphs encode information on this activity (through the numbers of neighbors of each ego), we group the ego-subgraphs into separate sets, each corresponding to a state: for each Egocentric Temporal Neighborhood collected at time $t \in s_e$, the resulting ego-subgraph is put into the set associated to s_e.

Overall, we thus end up with two different partitions of the temporal snapshots: one based on the Laplacian distance between snapshots, which will be used to adjust the overall group structure of the synthetic temporal snapshots created by our methodology, and one based on the variations of activity, which will determine which set of ego-subgraphs is used to create each surrogate temporal snapshot.

Reassembling the building blocks to generate a surrogate temporal network

Provisional snapshot. The surrogate network is progressively built one snapshot after the other, starting from time 1 to the final step T, with an iterative procedure. Each snapshot t is built making use of information on the previously built snapshots $t - d, \cdots, t - 1$, and of the information on the Laplacian state $s_\mathcal{L}$ and activity state s_e to which the snapshot t of the original network belongs.

Leaving to the Methods section the description of the initial creation of the first d snapshots, we present here the procedure to generate a generic temporal snapshot $t > d$, which is inspired by the configuration model for static networks[4]. While, in the configuration model, one assigns to each node

a desired degree, and one tries to match the different nodes in order to satisfy these requests, here we will assign to each node i in the snapshot t a desired set of neighbors, combining neighbors of i in the preceding snapshots and new neighbors, in order to eventually obtain a network with the same distribution of ego-subgraphs as the original network. Then, similarly to the configurational model, connections are built in order to match the requests of the various nodes.

More in details, to generate snapshot t of the surrogate, assuming that all the previous layers ($t' < t$) have already been generated, we consider all nodes one after the other and assign to each their interactions at time t based on the previous ones, as sketched in Fig. 1c. Specifically, for each generic node i we consider it as the ego-node, and we observe its neighborhood in the previous d surrogate snapshots $t - d, \cdots, t - 1$. This is an Egocentric Temporal Neighborhood of length d, that we map onto an ego-subgraph g_i^d (as shown in Fig. 1(c)). We then consider the set of all ego-subgraphs of length $d + 1$ of the original network associated to the state s_e to which t belongs. Among these, we select those whose first d time-steps coincide with the ego-subgraph g_i^d (in the example of Fig. 1c only the two highlighted ego-subgraphs satisfy this condition). Note that, although comparing graphs is in general a computationally heavy process[56], the fact that ego-subgraphs have been mapped onto binary strings makes this step very efficient, only requiring to find the binary strings of the set whose first d steps match the string describing g_i^d. The resulting subset of strings (or ego-subgraphs) determines the possible extensions of the neighborhood of node i in snapshot t that are compatible with the statistics of ego-subgraphs in s_e. We extract one of these possible extensions at random, with probability proportional to the observed frequency of the corresponding ego-subgraph. The extracted extension determines the desired neighborhood of i in snapshot t, which can include nodes already in interaction with i in the previous time snapshots, and interactions with new nodes. We represent the first case by directed desired links from i to the neighbors, and the second case by stubs, i.e., half links not yet connected to other nodes. In the example of Fig. 1c the random extraction yields an Egocentric Temporal Neighborhood of i on $[t - d, t]$ in which one of the neighbors at t was already neighbor at $t - 2$ and $t - 1$, while another neighbor is new, and is thus depicted as a stub (with a question mark). Once this process has been applied to every node, we thus obtain for the snapshot t a static "provisional" network with directed links and stubs, which encodes the desired neighborhoods of all nodes.

Link confirmation. In order to build the surrogate snapshot at time t, we now need to decide which of the desired neighborhood directed links of the provisional network to transform into actual links of the surrogate network, and how to connect the stubs. Indeed, not all the desired neighborhood links can be satisfied: for instance, if there is a directed link from i to j but not from j to i, it means that node i needs to connect to node j to have its desired Egocentric Temporal Neighborhood, but that a link between i and j in the snapshot t would frustrate the desired neighborhood of j by making its Egocentric Temporal Neighborhood different from the one extracted in the creation of the provisional network (vice-versa, not having the link between i and j would fit j's neighborhood but frustrate the one of i). To decide which desired neighborhoods to satisfy, we thus perform now the "confirmation" stage of the procedure, to go from the provisional snapshot to the actual surrogate one. The confirmation will depend on the Laplacian state $s_\mathcal{L}$ corresponding to time t in the original network.

First, we accept each reciprocal request (when both directed links $i \rightarrow j$ and $j \rightarrow i$ exist in the provisional network), transforming them into links of the surrogate snapshot, as the establishment of such links contributes to creating desired Egocentric Temporal Neighborhoods of both i and j. All the unidirectional links ($i \rightarrow j$ without reciprocal request) instead represent links that would satisfy one node and frustrate the other one, so we accept half of them[40]. The choice of the ones to accept is made taking into account (i) a mechanism of long term memory on the links, (ii) the average node clustering coefficient in the original snapshot, and (iii) the network's organization into groups at time t. More specifically, the confirmation process is designed to preferentially accept links:

- (i) that already appeared in previous snapshots of the surrogate network corresponding to the state $s_\mathcal{L}$, repeating interactions between nodes that have already met in that Laplacian state;
- (ii) that increase clustering, i.e., closing triangles by connecting nodes that have common neighbors;
- (iii) where the two nodes belong to the same group, according to the nodes partition in the state $s_\mathcal{L}$, with a probability that depends on the value χ of that state (i.e., if the original network has low/high modularity this effect is weak/strong), thus yielding a level of modularity in the generated snapshot similar to the one of the original snapshot.

In practice, we compute for each unidirectional link $i \rightarrow j$, the quantity

$$s_{ij} = (1 + \gamma_{ij})(1 + \rho_{ij}) \tag{3}$$

with γ_{ij} the number of common neighbors of i and j in snapshot t (counting all the possible bidirectional links to confirm) and ρ the memory matrix defined as follows: the element ρ_{ij} counts the number of past interactions between i and j in the surrogate that took place in snapshots up to $t-1$ corresponding to the same state $s_\mathcal{L}$ of time t, divided by the maximum value of ρ. Favoring pairs with larger γ_{ij} implies increasing clustering in the snapshot that we are generating, while favoring pairs with large ρ_{ij} corresponds to a (long-term) memory mechanism (a reinforcement process known to create heterogeneity between the aggregated strength of the connections[21,24]). We thus sort all the pairs in decreasing order of s_{ij}. Moreover we use the information about the groups to which each node belongs and the value χ of inter-groups versus intra-groups connections (found in the original network for the state $s_\mathcal{L}$ corresponding to time t) as follows: we review the list of unidirectional links, starting from the pair with the largest s_{ij}, and accept each link with probability $1 - \chi$ if i and j belong to the same group, and with probability χ otherwise. We stop the process when we have accepted the needed number of pairs (half of the total). If the end of the list is reached before this number is reached, we choose the missing links at random in the list. We note that this method is designed to reach large values of the clustering, since one accepts preferentially the links with large values of the number of common neighbors. It is however possible to modulate this effect and obtain lower values of the clustering by multiplying γ_{ij} by a factor $c < 1$ in Eq. (3).

Finally, we couple the stubs (deleting one if we have an odd number of them) using a similar process. We first create the list of all the nodes with stubs, each node being repeated a number of times equal to its number of stubs. We choose a random node i from this list, and compute s_{ij} for all the other elements $j \neq i$ of the list. We sort the nodes j in decreasing order of s_{ij} and, as above, starting from the largest value, accept the match between i and j's stubs with probability $1 - \chi$ or χ, depending whether i and j belong or not to the same group. As soon as a match is accepted, we create the corresponding link (i, j) in the surrogate snapshot (if not yet existing), and remove one copy of i and j in the list of nodes. We repeat the process until all the stubs have been coupled.

At the end of the confirmation stage, we obtain a static network that becomes the snapshot t of the surrogate network. We iterate to create snapshot $t + 1$ and so on, until we reach the final time T.

Since we are using the memory of past interactions to generate each sapshot, we do not consider valid the first T_ρ snapshots that are generated, where memory is not significant yet, so we generate $T + T_\rho$ snapshots in total and the surrogate network will correspond to just the last T. The parameter T_ρ is here arbitrarily set to correspond to 24 hours.

Note that the surrogate generation process is stochastic, as both the choice of the desired Egocentric Temporal Neighborhoods of the nodes and the confirmation steps include random choices. Using different realizations of these choices thus allows us to produce as many different surrogate temporal networks as needed, all reflecting the same statistical properties of the original data. Since the procedure takes into account the statistics of the ego-subgraphs (E), the structural correlations (S) like modularity and clustering, and the long term temporal correlations (T), in the following we

will refer to it as the EST model. We will consider also two simplified versions: the ES model (without temporal correlations, i.e., setting $\rho_{ij} = 0 \ \forall \ i, j$) and the E model (without temporal nor structural temporal correlations, i.e., $\rho_{ij} = 0$ and $\gamma_{ij} = 0 \ \forall \ i, j$).

Temporal extension. The procedure described above generates surrogate networks with the same temporal length T as the original network. We note however that it is possible to iterate the procedure beyond T and to create as many additional snapshots as desired, and thus to generate surrogate networks of arbitrary length. To do this, we need however to associate each time $t > T$ to a pair of states s_e and $s_\mathcal{L}$, as the original network does not contain such information. One possibility is to use the set of known states, with (for instance) a periodicity depending on the context (for the primary school data for instance a periodicity of 24 h can be suitable to reconstruct the general temporal evolution of successive days). Such a procedure allows to extend datasets without repeating identical interactions, but mimicking the general behavior of the original datasets to generate new synthetic stochastic interactions.

Empirical evaluation
We illustrate the procedure using eight temporal networks of social interactions as original networks. Five of them have been collected by the SocioPatterns collaboration (www.sociopatterns.org) and report face-to-face contacts between sets of anonymized individuals, collected using wearable sensors in diverse environments, including schools, workplaces and conferences, and have a temporal resolution from 20 s. The other three datasets have been collected as colocation data in an elementary school, a middle school[47] and a university campus[48] with a resolution of 20 s for the first two and 5 min for the last. We show in the main manuscript the results when using the primary school contact data of Sociopatterns[43] as original network, and show the results for the other datasets in the Supplementary Information (Supplemetary Note 6). Indeed, this dataset, which describes the interactions between students and teachers during two days, entails rich intertwined structural and temporal features such as groups of nodes (classes) that are densely connected during some periods (classes) and form larger, less dense clusters during other periods (lunch breaks).

Using this dataset as original network (with here a temporal resolution of 5 min) we apply our stochastic procedure (with $d = 2$) 10 times, thus obtaining 10 different surrogate networks, for which we report the averaged results. Note that we report the results both for the most refined version of the procedure (EST) and for the two baselines E and ES, in order to understand which aspects of the procedure are most needed to best mimic the various features of the original network. We first compare the surrogate and the original networks under the lens of several structural and temporal features of temporal networks, and we then study whether several dynamical processes unfold in a similar fashion on the surrogate and on the original data.

The basic method that only uses Egocentric Temporal Neighborhoods and ego-subgraphs is very fast (as shown in ref. 40 for a method similar to our E process). It is slightly slowed down by the additional structural and temporal features but the implementation time remains short: it takes from 20 s to 10 min on a standard laptop to generate one surrogate network of those shown in main text and Supplementary with the EST method (the most complicated). The time variability is mainly due to the number of nodes, that for the networks we tested go from 113 to 675.

Structural and temporal properties. All three procedures (E, ES and EST) produce surrogate temporal networks whose number of interactions per temporal snapshot mimics well the original one (top panel of Fig. 2a), albeit with smoothed out fluctuations. This is expected from the method's design, as the original temporal activity has been divided into states s_e, which are reproduced in the procedure. Similar results were obtained by ref. 40, whose method corresponds to the E procedure with a temporal partition built on an arbitrary time-scale of one hour, instead of being extracted from the data through the states s_e.
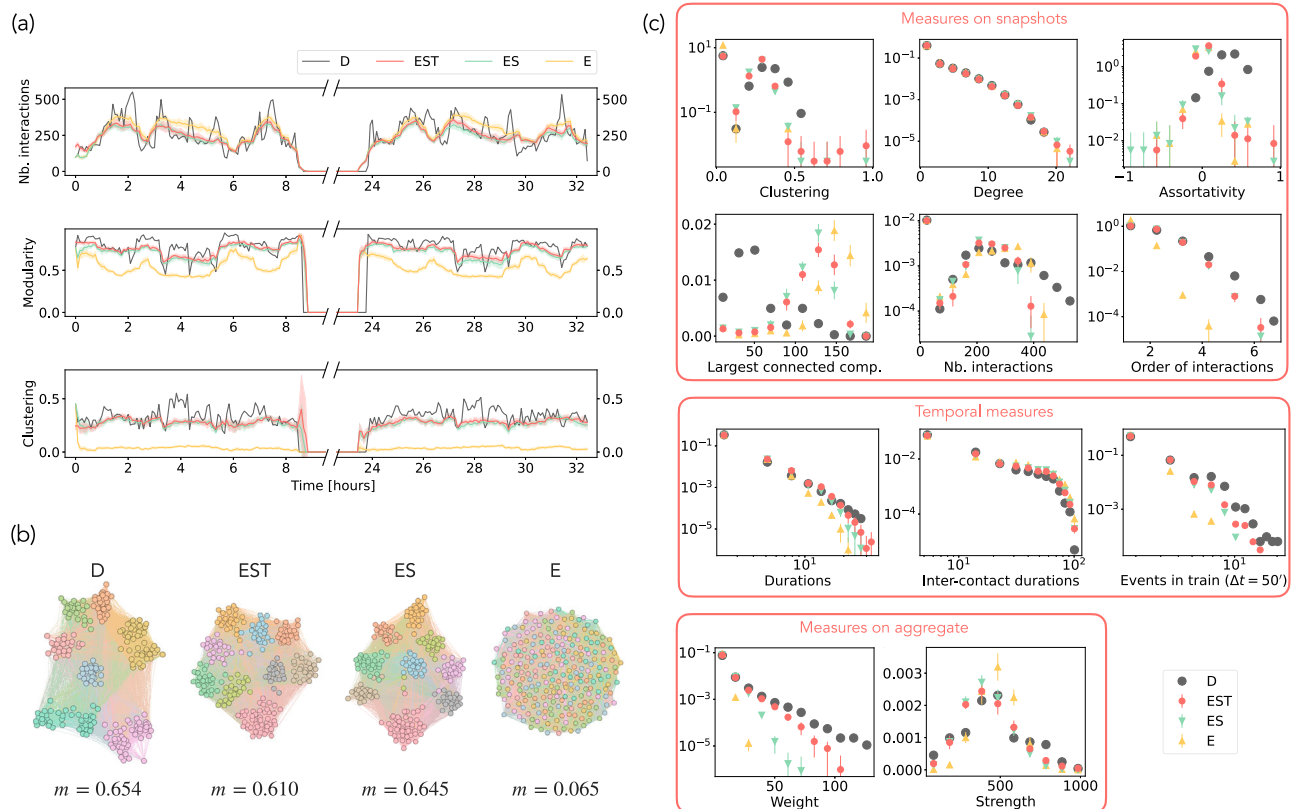
Fig. 2 | Structural and temporal measures. a Number of interactions, modularity and clustering for each temporal snapshot of original network (D), here the primary school dataset, and for surrogate networks obtained by the EST, ES, and E methods (average over 10 realizations for each method; the shaded area represents the standard deviation). **b** Aggregated networks for original and surrogates (one realization for each method), where nodes colors reflect the nodes partition obtained by the Louvain algorithm and m is the corresponding modularity index. **c** Distributions of values for several structural and temporal quantities, divided in three groups according to the measure type. For the surrogate networks we average over 10 realizations (vertical bars give the standard deviation).

The lower panels of Fig. 2a report the evolution of more interesting quantities, namely the average clustering and the modularity of each snapshot. The E model, which does not take into account structural correlations, yields snapshots with very low clustering and lower modularity than the original ones. In contrast, both the ES and EST procedures produce values similar to the ones of the original data. Panel (b) of Fig. 2 gives an illustration of this point by showing the networks resulting from the temporal aggregation of the original and surrogate temporal networks (one realization for each surrogate method) where the nodes have been colored according to their partition in communities obtained with the Louvain method[55]. The panel also gives the value of the modularity m of these aggregated networks, averaged over 10 realizations of the surrogates (modularity value obtained in each realization through the Louvain method). Both ES and EST methods yield structures comparable to the original one, while the E method produces a rather homogeneous network with no group structure.

Panel (c) of Fig. 2 gives a more systematic comparison of the properties of surrogates and original dataset by showing the distributions of eleven quantities characterizing the statistics of individual snapshots, temporal statistics and aggregated ones.

Instantaneous snapshot properties. The first six quantities of Fig. 2c are defined for static networks: we compute them on each single snapshot and display the resulting distributions. They consist in: clustering, degree, assortativity, size of largest connected component, number of interactions and orders of interactions. The *clustering* is computed as $3\frac{\#triangles}{\#triads}$ where *#triangles* is the number of closed triangles (triplets i, j, k such that all links $i - j, j - k, k - i$ exist in the snapshot) and *#triads* is the number of triads in which at least two of the links exist. The *degree* refers to the degree of each

node in each snapshot (instantaneous number of neighbors). The *assortativity coefficient* measures the correlation between degrees of connected nodes in each temporal snapshot[57]. The *largest connected component* of a temporal snapshot is the largest subgraph that is connected in that snapshot. The *number of interactions* refers to the number of links in a snapshot. The *order of interactions* is defined by promoting all the maximal cliques (groups of all-to-all connected nodes) in each temporal snapshot to higher-order interactions[58]; the order of an interaction is then given by the number of nodes in a clique minus one (first order interactions involve two nodes, second order involve three, and so on).

In most cases, the distributions computed on the surrogates obtained with the EST procedure are very similar to the original ones. Some properties (distributions of degree and of the number of interactions) are actually well reproduced even with the simplest model E. On the other hand, the E procedure yields small clustering values and small orders of interactions, while taking into account structural correlations in the procedure (Eq. (3)) yields distributions close to the original ones. Note that the generation method considers only pairwise interactions, does not build higher-order interactions (involving more than two nodes) and does not rely on higher-order concepts. Nevertheless, the combination of clustering and community structure effects increases the probability to generate cliques in the surrogate snapshots, and allows us here to reproduce the statistics of instantaneous cliques, which can be interpreted in the present context as higher-order interactions[58,59]. Two of the distributions are not well reproduced by the surrogates, highlighting some limitations of the model. First, the assortativity is always around zero, independently on the degree correlations in the original network, which are instead slightly positive in all the considered datasets (see Supplementary Note 6). In the generation procedure indeed there is no mechanism that takes into account the neighbors' degree when
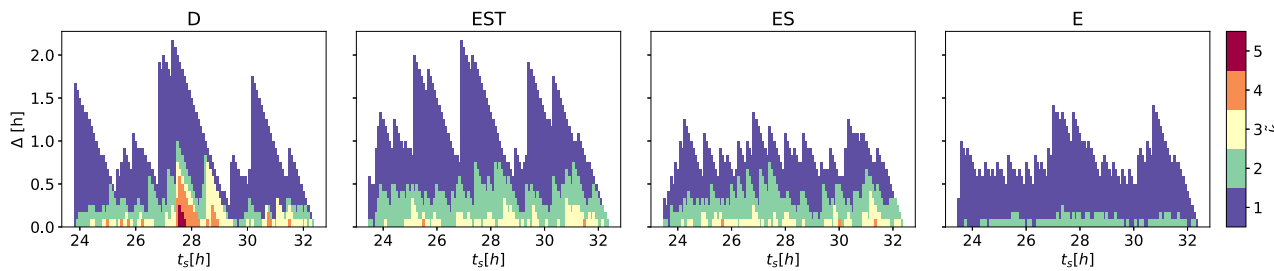
Fig. 3 | Span-core decomposition. For different $\tilde{k}$ (different colors) we show the duration of span-cores starting at various times t_s of the second day of the primary school network and of its surrogates.

confirming the links or connecting the stubs, hence no correlation between degrees emerges. Second, the distribution of the sizes of largest connected components is shifted to larger values with respect to the original network. This is true for all the datasets considered (see Supplementary Note 6), and this effect is particularly large for the primary school dataset shown in Fig. 2. The reason can be ascribed to the rigid constraints of a school context, which are not directly implemented in the surrogate but only loosely reproduced through the division in temporal states: during lectures at school, the children are separated into classes and there are no interactions between individuals of different classes, so that the largest connected components in the corresponding snapshots are small. In the surrogates instead, this is not strictly implemented: the generation of a group structure, even with a high modularity value, still leaves the possibility for some interactions between groups, which strongly increase the connected components sizes.

Temporal properties. The second group of plots in Fig. 2c displays the distributions of three important measures characterizing the temporal evolution of interactions in a temporal network, and measured for all pairs of interacting nodes and over the whole time span of the networks. The *duration* of an interaction between two nodes i and j is defined as the number of consecutive snapshots in which i and j are connected. The *inter-contact duration* is instead the number of consecutive snapshots between two successive interactions of i and j (excluding the empty snapshots such as those representing the nights). The *burstiness parameter*[60,61] quantifies the heterogeneity of the inter-contact durations distribution as $B = \frac{\sqrt{n+1}r - \sqrt{n-1}}{(\sqrt{n+1}-2)r + \sqrt{n-1}}$ with r the ratio between standard deviation and mean inter-contact duration, and n the sample size (number of inter-contact durations used to create the distribution). The burstiness is -1 for a periodic time series, 0 for a Poisson one, and it is positive for a bursty distribution. Finally, we consider the distribution of the *number of events in a train*[19]: an "event" is defined here as an uninterrupted interaction between two nodes (with its duration), and a "train of events" as a series of consecutive interactions between the same two nodes such that the interval between the end of an event and the beginning of the successive one is smaller than a parameter Δt. Multiple trains of events can exist for each pair of nodes and the desired measure is obtained by counting the number of events in each train for each pair of nodes. A broad distribution of this quantity unveils the presence of temporal correlations in the network[19].

Figure 2c shows that the EST procedure yields surrogate networks able to reproduce well the statistics of these three quantities, while the E and ES procedures yield narrower distributions of the interaction durations and of the numbers of events in a train. The burstiness of the EST surrogate (averaged over 10 realizations) is also closer to the original one ($B_{data} = 0.33$, $\langle B \rangle_{EST} = 0.28$, $\langle B \rangle_{ES} = 0.26$, and $\langle B \rangle_{E} = 0.16$) even if all three methods yield comparable inter-contact durations distributions. Overall, the EST surrogate thus reproduces well the temporal heterogeneities and correlations observed in the original empirical dataset.

Aggregated network. The last two quantities displayed in Fig. 2c are measured on the networks aggregated on their whole temporal span: we show in

the bottom panel the distributions of the links *weights* and of the node *strengths*: the weight of a link is equal to the number of snapshots in which the link has been active, while the strength of a node is the sum of the numbers of instantaneous neighbors of this node in each snapshot. For the E procedure, the links weights display a narrow distribution, reaching only small weight values. This is due to the fact that this procedure does not entail structural nor memory effects, so that the new connections of a node created in the generation and confirmation of a snapshot are taken at random among all possible ones, and are less repeated than in the ES and EST cases. Taking into account structural correlations leads to a slightly broader distribution but, as expected from previous studies[21,24], memory effects are needed to obtain distributions more similar to the original ones.

Temporal network structures. Several characterization tools have been developed recently to analyze the complex interplay and correlations between topological and temporal aspects of temporal networks: we consider here two of these tools, which highlight the existence of mesoscale structures in temporal networks.

We first perform the span-core decomposition[16] of the original and surrogate networks: it decomposes a temporal network into hierarchies of subgraphs of controlled duration and increasing connectivity, generalizing the core-decomposition of static graphs. Specifically, a span-core g of order $\tilde{k}$ is defined on an interval of Δ consecutive timestamps, such that all nodes in g have at all timestamps of that interval at least $\tilde{k}$ stable neighbors in g (i.e., the links to these nodes are present during all timestamps of the interval). Highly connected (large $\tilde{k}$) and stable (large Δ) span-cores have been shown to be structures relevant in spreading processes[62], and empirical temporal networks are often characterized by the presence of such structures, which are absent from random, uncorrelated temporal networks[16]. We report in Fig. 3 the result of the span-core decomposition for the original dataset and for one instance of each surrogate, by showing the duration and the connectivty of the span-cores starting at various times. Span-cores of durations comparable to the ones of the original data are obtained with the EST method, while the E and ES methods, which lack long-term memory or structural effects or both, yield less rich structures.

We then investigate the temporal rich club phenomenon[17], i.e., the tendency of nodes that are well-connected in the aggregated network to form structures that are simultaneous and stable in a temporal network. To measure this tendency, we consider, for the original data and for the surrogates, the subset of nodes $S_{>\bar{k}}$ of nodes with degree larger or equal to $\bar{k}$ in the aggregated network, and we compute $\varepsilon_{>\bar{k}}(t, \Delta)$, defined for each time t as the fraction of links that connect the nodes of $S_{>\bar{k}}$ in a stable way from t to $t + \Delta$. The maximum $M(\bar{k}, \Delta)$ over t of $\varepsilon_{>\bar{k}}(t, \Delta)$ (the Temporal Rich Club coefficient[17]), shown in Fig. 4 as a function of $\bar{k}$ and Δ, quantifies whether the connections between the nodes highly connected in the aggregated network were simultaneous, dense and stable. Connected and stable structures are observed in the dataset, with a temporal rich club effect of increasing density with increasing $\bar{k}$, and also in the surrogate produced by the EST method, while the effect is much weaker for the ES case and totally absent in the E model.

These results show that our methodology, when taking into account both structural and temporal characteristics of the original network, is able
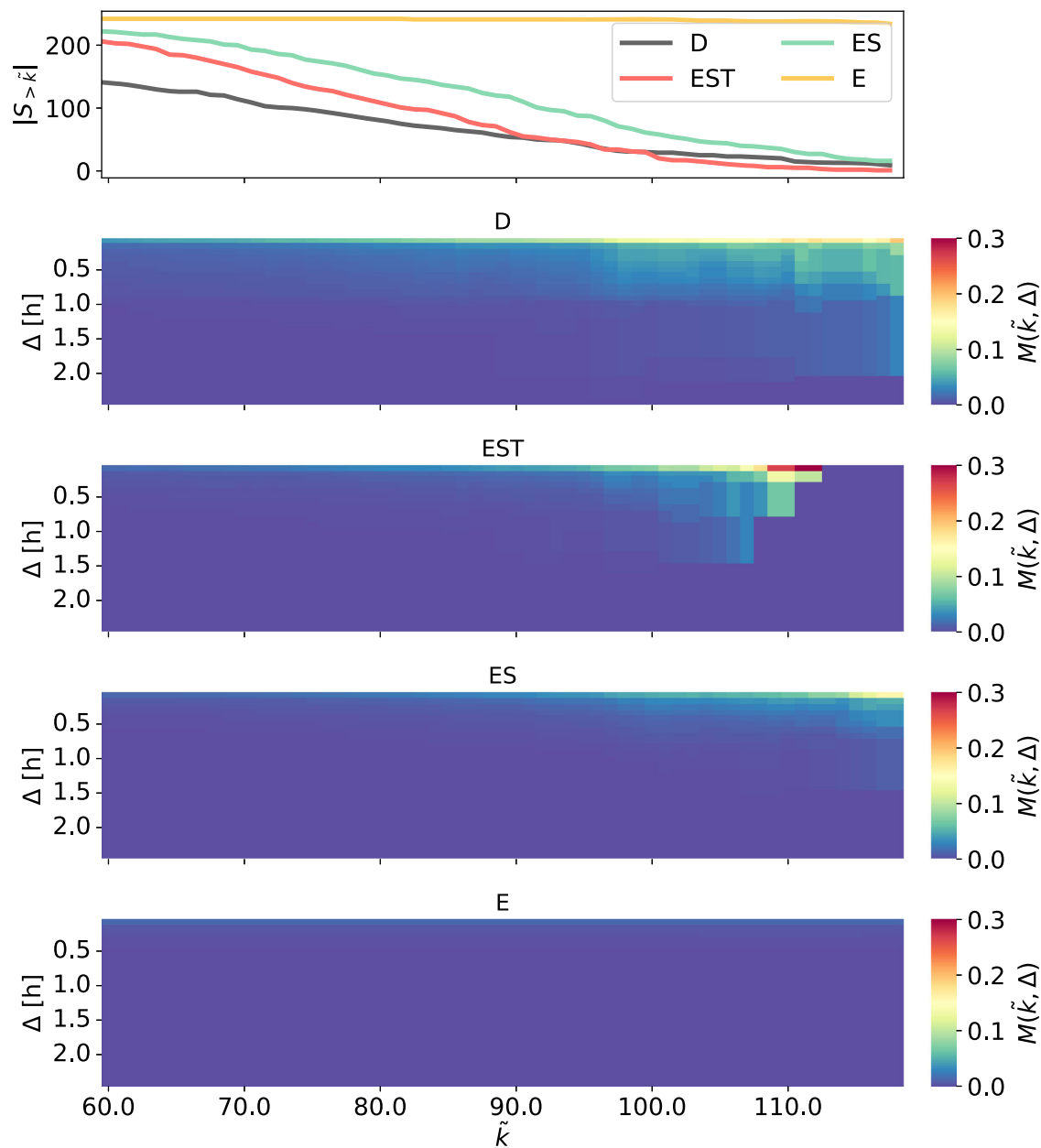
Fig. 4 | Temporal rich club. Size of the subset $S_{>\bar{k}}$ (top panel) and maximum $M(\bar{k}, \Delta)$ over time of $\varepsilon_{>\bar{k}}(t, \Delta)$, as a function of $\bar{k}$, for several values of Δ (lower panels).

to generate surrogate data that entails highly complex structures observed in the empirical temporal networks.

Dynamical processes. As discussed in the introduction, surrogate networks are particularly important as a substitute to empirical data to perform numerical simulations of a variety of dynamical processes. As the outcomes of these processes depend on numerous properties of the network they take place on, we investigate here whether the processes simulated on the original and surrogate data unfold in a similar manner. We choose three paradigmatic processes that have been largely studied on static and temporal networks[1]: (i) a model describing the spreading of a schematic disease[8], (ii) a model for opinion dynamics with bounded confidence, i.e., such that agents can change opinion by interacting with other agents whose opinion is not too different[49], and (iii) a model for the emergence of conventions in a population[52].

Spreading dynamics. The first dynamics that we consider is the Susceptible-Infected-Recovered (SIR) model for the spread of infectious diseases, where

the nodes of the network can only be in one of the three states: Susceptible nodes (S state) have not yet been reached by the disease, but can be infected with probability β per unit time (per snapshot) when interacting with an infectious node (in I state)[8]. Infectious nodes then recover spontaneously with probability μ per unit time, entering the R state. Each simulation starts with one initial seed, represented by a node chosen uniformly at random, which is put in the I state at a random time t_0, while all the others are susceptible. The simulation ends when no node is in the I state any more (all nodes have either been infected and then recovered, or have remained in the S state). If the last snapshot (time T) of the temporal network is reached with the process still active, we repeat the temporal network starting from the first snapshot.

We perfom simulations of the process for varying β and μ and using the original and surrogate datasets. For each simulation we compute the basic reproduction number R_0 and the final number of recovered R_∞, which quantify respectively how many other nodes the initial seed infects, and the final epidemic size, i.e., the size of the network that has been reached by the spread (see the Methods section). The heatmaps in Fig. 5 display the mean
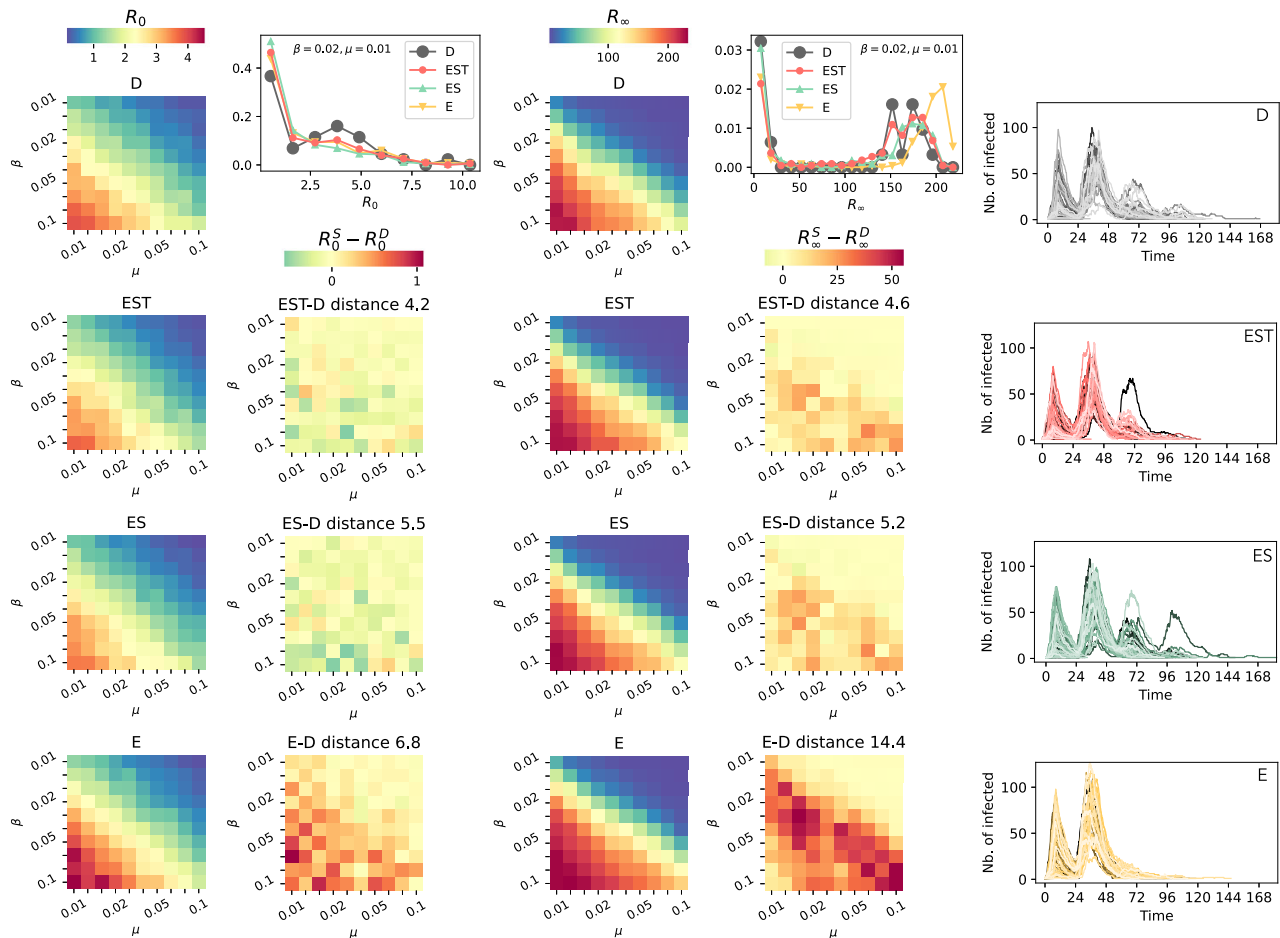
Fig. 5 | SIR model. The heatmaps of the first column show R_0 at varying β and μ for original and surrogate networks (averaged over 200 simulations for each pair of parameters). The second column shows the differences between the values obtained by simulations on surrogate networks and the ones obtained with the original network. On top of these heatmaps we report the Canberra distances[66] between the two matrices of values. We also report on top of the column the distributions of R_0 values for a specific choice of β and μ. Analogously, the heatmaps on the third column show the mean values of R_∞ for all the networks and those of the fourth column the differences between surrogates and original network. The plot above reports the distribution of R_∞ for a given pair of parameters β and μ. With the same choice of parameters, the plots on the right show the time evolution of infected individuals on all the networks for 75 simulations. For these specific plots, to facilitate the comparison, the start of the simulations is fixed at $t_0 = 0$ for all the realizations.

values of R_0 for the different networks (first column) and the difference between the mean values obtained in simulations on surrogates and dataset (second column), for each set of parameters (β, μ). The third and fourth columns show heatmaps for the average values of R_∞ and for the differences in these average values obtained in the simulations on surrogates and on the original data. The plots on top of the second and fourth columns show, for a specific choice of β and μ, the distributions of R_0 and R_∞ for each underlying network. Finally, the rightmost column shows the temporal evolution of the number of infectious individuals for 75 simulations starting all at $t_0 = 0$ performed on the original network and on each surrogate.

The values, variation patterns, and distributions of R_0 and R_∞ obtained with the EST model are very similar to the ones obtained for simulations run on the original dataset. The temporal variations of the number of infectious is also well reproduced. A similar picture is obtained when simulations are performed on surrogate networks obtained with the ES model. In surrogates produced by the E model, the lack of structure leads to a faster process with larger impact[8]: more nodes are infected at the beginning (larger R_0) and the spreading is faster (epidemic curves spanning typically shorter timescales), reaching larger parts of the network (larger R_∞). In contrast, the group structure present in the ES and EST surrogates has the effect, as in the original dataset, to limit the possibility of spreading between groups. Moreover, the memory effects, which leads to the repetition of interactions, also impact the spread[12]. Overall, the better reproduction by the EST of the

structural and temporal features highlighted in Fig. 2 leads also to outcomes closer to the one of the original data[10].

Deffuant model for opinion dynamics. The second dynamical process that we consider is the Deffuant model of opinion dynamics[49,50]. In this model each node i represents an individual, endowed with an opinion represented by a real valued variable x_i between 0 and 1. The initial state is given by assigning an opinion extracted uniformly at random to each individual. Then, at each snapshot t, opinions of pairs of individuals who interact on the network can change, if and only if their opinions differ less than a fixed parameter q, i.e., $|x_i(t) - x_j(t)| < q$ (a rule known as "bounded confidence", to express the concept that individuals tend to exchange only with other individuals whose views do not differ too much). If this condition is met, the two individuals update their opinion to a common middle ground:

$$\begin{aligned}
x_i(t+1) &= \frac{1}{2}\left(x_i(t) + x_j(t)\right) \\
x_j(t+1) &= \frac{1}{2}\left(x_i(t) + x_j(t)\right).
\end{aligned} \tag{4}$$

The dynamics evolves until the opinions of all interacting pairs cannot change anymore (either because they are already aligned or because they are farther apart than q). See Methods section "Deffuant model and Naming

Game on temporal networks" for more details. Such a process tends to align opinions of interacting individuals but does not necessary lead to a globally aligned population because of the bounded confidence mechanism: even in a population where all individuals can potentially interact with each other, small values of q lead to the separation into groups of individuals who share the same opinion, but such that the opinion of different groups differ more than q, making communication between groups impossible[49].

In the case of a process taking place on a (temporal) network, the situation is a bit more complex. The final state corresponds to a separation of the population into groups of nodes inside which all nodes share the same opinion, such that the opinion in a group differs of more than q from the opinion of the other groups *with which it has interactions*. However, there can potentially be two groups of individuals whose opinions differ of less than q, but who never interact directly along the network. We thus characterize the final state by the number of subnetworks of individuals with a homogeneous opinion, which depends on q and on the interplay between the dynamics of opinions and the network's structural and temporal properties. To obtain this number in practice, we consider the aggregated network, and remove all the links connecting nodes whose opinions differ more than q in the final state: we then count the number of remaining connected components (which constitute the opinion subnetworks generated by the process).

We report in Fig. 6a the final number of such opinion groups generated when simulating the Deffuant model on the original and surrogate networks, as a function of q. Panel (b) reports moreover the convergence time of the process, i.e., the number of snapshots after which the opinions do not evolve any more. Interestingly, the final number of opinion groups is well reproduced even with the simulations on the surrogate networks obtained with the simplest E method. However, the process is there much faster and the addition of structure in the ES and EST surrogate methods leads to dynamics much closer to the one on the original network.

Naming Game. Lastly, we consider the Naming game, a process where the nodes representing individuals aim to reach a consensus on the name to give to some object or concept[51,52]. For simplicity we restrict the possibilities to only two names, A and B. Each node has an inventory of possible names which at the beginning contains either A or B (chosen at random). The simulation starts at a random temporal snapshot of the network. At each time, for each interacting pair of nodes, we choose randomly one node to act as speaker and one as hearer. The speaker chooses a random name in its inventory and proposes it to the hearer. If the hearer does not have it in its inventory, the name is added to the inventory. If instead the name was already present in the hearer's inventory, the two nodes agree on this name by removing the other possible name. This happens with probability η, representing the propensity of the hearer to accept the name[51]. The process ends when all the nodes agree on the same name (the other name has then disappeared from all nodes' inventories). See Methods section "Deffuant model and Naming Game on temporal networks" for more details. Figure 7 reports the distributions of convergence times for two values of η and for Naming Game processes simulated on the original data and on surrogate networks. It also shows, for several realizations of the process, the temporal

evolution of the number of nodes that have the finally winning name in their inventory.

We observe here a similar phenomenon as in the Deffuant model: simulations performed on the surrogate network produced by the E method have a typically much faster convergence than for the original case. The ES and EST methods produce surrogate networks that better reproduce complex features of the original data, leading to dynamics more similar to the original one, even if the convergence time remain slightly underevaluated (note that the peaks in the distribution of convergence times for runs on the primary school data correspond to the successive day/night sequence: the process can converge either in the first day, in the second day or in few cases need even an additional simulation day).

Robustness. The Egocentric Temporal Neighborhood strategy is based on the idea that the behavior of a node is strictly related to the immediately previous interactions, where the time scale is set by the parameter d. In the examples that we have shown, d was fixed to 2 but values 3 and 4 have been tested and are shown in Supplementary Note 1. The results with $d = 3$ are similar to those obtained with $d = 2$ (but slightly worse on some important measures like the degree), while those with $d = 4$ turn out to be significantly worse. In fact, increasing d allows to retain more information about last contacts, but, since many more ego-subgraphs are possible, the frequency of each one is reduced and the results are affected by limited size effect: it is possible that a node in the surrogate snapshots $[t - d; t - 1]$ explores a set of interactions that does not match any of the ego-subgraphs of length $d + 1$ of the original network. In that case neighbors are progressively removed until a match is found. In other words, increasing d can result in overfitting: trying to reproduce ego-subgraphs with more information paradoxically leads to a surrogate whose ego-subgraphs are less conform to those of the original network.

Another robustness investigation is related to the temporal decomposition into states. As discussed in ref. 15, the decomposition can depend on the measure used to compute the distance between temporal snapshots, and also on the clustering algorithm considered (and on the choice of the number of states). However, we show in Supplemenatry Note 2 that the properties of the surrogate temporal networks remain similar when using a division between states with an arbitrary time scale of one hour, hinting at a robustness of the procedure with respect to the choice of the division into temporal states.

Finally, we test the possibility to generate a surrogate of the full temporal extension of the original network even if only the initial part of the network is known. To this aim, we assume to have access only to the first snapshots of the empirical data (corresponding to the first 24 h in most cases, or to the first week for longer datasets), representing between 14% and 75% of its total length. As described in the Supplementary Note 7, we then generate surrogates covering the full temporal length of the original datasets. Results are shown in Supplementary Note 7 and highlight the similarity of these surrogates with the original networks: No significant difference is observed with the case of surrogates obtained by using the information collected on the complete original datasets. Note that we also show in the Supplementary Note 4 an example of temporal extension beyond the time

Fig. 6 | Deffuant opinion model. Number of disconnected opinion groups (**a**) and convergence time (**b**) at varying q, averaged over 200 simulations on each network (for each surrogate method, we perform 10 realizations of the surrogate network and 20 runs of the opinion model on each realization). The curves show mean and standard deviation (shaded area).
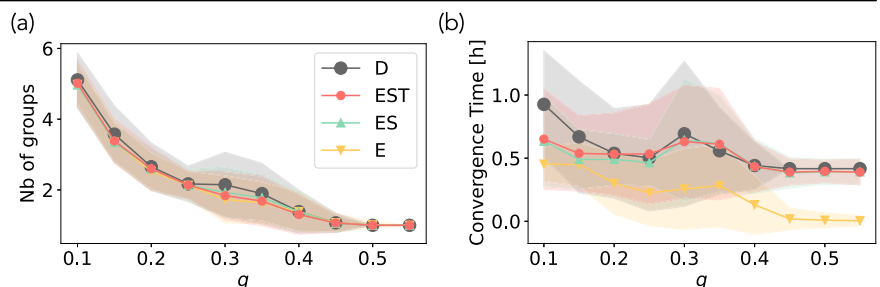
Fig. 7 | Naming game. The top panels show the distribution of the convergence time for the Naming Game simulated on all the networks (100 simulations on the original network and 10 simulations on each of the 10 realizations of the surrogate networks) with $\eta = 0.4$ and $\eta = 0.9$. The four plots below show the time evolution of the fraction of nodes that have in their inventory only the name that will eventually win. For each value of η 10 realizations are displayed for each network.
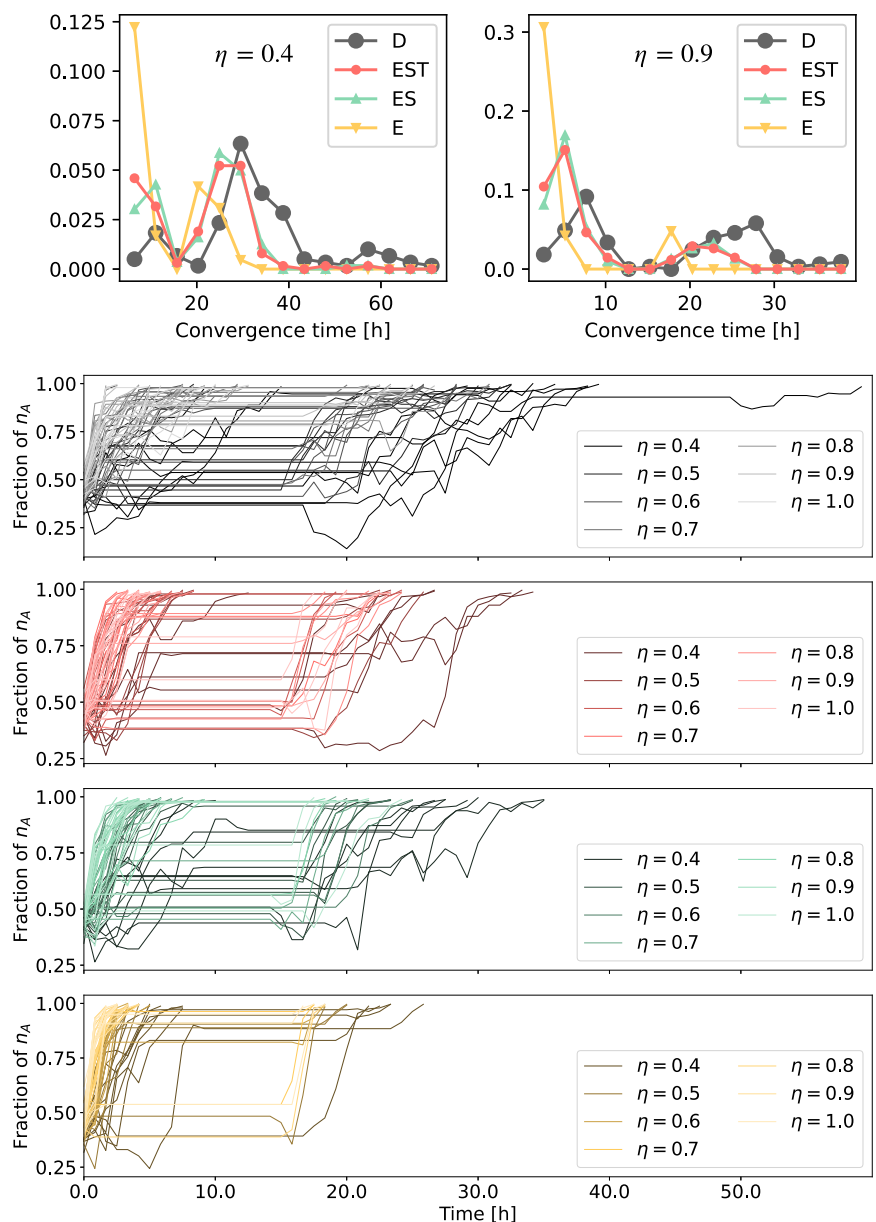
span of the original dataset, generating surrogates for the primary school with a temporal length of 40 days, while the original dataset is only 2 days long.

Discussion

The method that we have proposed here to generate surrogate temporal networks starts from the analysis of an original network, supposed to be known and that the surrogate data should mimic, i.e., the surrogate network, without being identical to the original one, should have similar statistical properties and structures. This analysis characterizes both its local and global properties and their temporal evolution. The developed methodology leverages the temporal division of the temporal network into states on the one hand, and in collections of ego-subgraphs on the other hand, to generate surrogate temporal snapshots one after the other by gluing together different ego-subgraphs, implementing in the process a long-term memory mechanism and mimicking the mesoscale organization of the original snapshots (clustering, organization in groups). We have shown that the resulting surrogate networks display a complex interplay of structural and temporal properties similar to the one of the original network. Moreover, we

have shown that simulations of a variety of dynamical processes on the surrogate networks yield outcomes similar to the ones obtained by simulations on the empirical data, even if some discrepancies can be observed, as surrogate data can never fully capture all the complex correlations of an empirical dataset[10,36].

In particular, it is important to note that the three dynamical processes that we have considered represent a broad variety of testing frameworks for our networks, as they differ in several fundamental aspects. For instance, in disease spreading the nodes states are discrete (S, I, R) and each node can only follow an irreversible process from S to I to R, without going back. Nodes having reached the R state do not take part any more in the process. In the Naming Game model the states are also discrete (A, B, AB) but each node can in principle pass from one state to another one an infinite number of times. In the Deffuant model, the states (opinions) are continuous ($x \in [0, 1]$) and also here there is a priori no limit to the number of times that a node can change opinion. The global outcomes that result from these process properties are also different and highlight different characteristics of the network: in the SIR model a fraction of the nodes is infected and then recover, in the Deffuant model the nodes are partitioned into opinions

clusters, and in the Naming Game groups of nodes having converged on one of the names can emerge, before a global convergence is obtained. The size of the fraction of the network that is affected by the SIR process, the number of opinion clusters in the Deffuant model, and the groups created in the Naming Game are impacted by the properties of the network of interactions in different ways[1,7]. For instance, memory effects impact the Deffuant model and the Naming Game, as repeated interactions between a pair of nodes allow them to maintain their agreement even if they also interact with other neighbors. In the Naming Game moreover, convergence between a node having only name *A* and one having only name *B* requires at least two interactions. The SIR model is less impacted by memory, as, once a link has been used to transmit the infection, it cannot be used for transmission any more. The SIR process is instead more affected by the burstiness of a temporal network: during a long time without interaction, a node cannot transmit but can recover spontaneously. On the other hand, its opinion or inventory in the other models simply does not evolve. Clustering and group structures also favor local spreading and opinion convergence, but can make it harder for the SIR process to go beyond the group it started from, or for different groups to converge on a common opinion (this last situation is clearly observed with one of the datasets corresponding to contacts collected in a high school, where the group structure is very strong, as described in Supplementary Notes 5 and 6).

The fact that simulations of such various processes on the surrogate networks provide a similar phenomenology as on the original data highlights the versatility of our method. We also note that, while we have here focused the presentation of our results on the ability of our method to generate surrogates similar to the original networks, an interesting outcome of our work also consists in the possibility to create synthetic datasets mimicking the observed ones but with specific network features altered and customized. Surrogate networks can indeed be generated with a different grouping in communities (e.g., a different number or different sizes of classes for a school), with a different organization of timescales (such as a rearrangement of activities timetable), or, as shown in the previous section, a longer total duration. In fact, an important use of surrogate networks consists in providing substrates with realistic properties on which dynamical processes can be studied on various enough time scales, even if long enough datasets are not available. An important example is given by simulations of realistic spreading processes, which often have longer timescales than most available datasets. It is then typically needed to use multiple repetitions of the same temporal network, which has consequences on the variability of interactions and hence on the realism of the observed behaviors[11,12,63,64]. The methodology presented here makes it possible to circumvent this difficulty by generating surrogate data of the needed length, avoiding to repeat exactly the same patterns[12].

The methodology has also some limitations worth discussing. First, as the surrogate networks are each based on an original dataset that they mimic, a large quantity of information about the data is needed. We have here assumed to have full knowledge of the original dataset and that it does not suffer from incompleteness. The case of incomplete data[10], where the population of nodes is only partially observable and hence the Egocentric Temporal Neighborhoods might be different from those of the complete dataset, is an interesting avenue for future work. Second, we have observed that, even if the method is general and can be applied to any kind of temporal network, we obtain better performances on denser networks. In fact the introduction of a preferential link confirmation by Eq. (3) is more effective if the amount of possible links to choose from at each time step is large enough: if instead only a few possible links are available, the variability is limited and so are the chances to find suitable links that correspond to high values of s_{ij}. For temporal networks with too diluted snapshots, we have thus considered lower temporal resolutions by partially aggregating the empirical data snapshots (e.g., on 5 min temporal windows for the primary school dataset, 20 min for the high schools and the conference, and one hour for the workplace dataset).

The methodology we have presented has focused on mimicking the interaction behavior of nodes at short time scales, as well as the instantaneous mesoscale structure and long-term memory effects. These ingredients have shown to be enough to reproduce a broad range of complex features of temporal networks. The versatility of the method makes it however possible to introduce additional mechanisms to reproduce other properties that could be unveiled by future studies of temporal networks. For instance, the confirmation stage could be tailored by changing the definition of s_{ij} to make it depend on other features of the nodes or pairs of nodes (e.g., nodes attributes). One example of additional mechanism is shown in Supplementary Note 5 where, instead of the simple modularity, a hierarchical clustering between communities is reproduced in the surrogates.

A further future development of the present work would be to use an existing dataset and produce realistic surrogate data with larger population sizes. This would yield the important benefit of being able to simulate dynamical processes on networks of large size without needing to actually collect the corresponding data (hence, for instance, without privacy concerns). Such a development is however far from trivial as the way in which the Egocentric Temporal Neighborhoods change when the population size increases has not yet been investigated, for instance. We therefore plan to tackle this point in future work.

Methods
Modularity and value χ
One of the steps of the original network analysis implies to separate the temporal snapshots into states $s_{\mathcal{L}}$ and to assess their group structure. Each state consists of several snapshots, i.e., several static networks. To find the best partition of nodes into groups we use the Louvain method[55] on the aggregate network of each state. Different states will hence correspond to different partitions (for example for the primary school data, the temporal state corresponding to the lectures periods yields a partition given by the classes, and the partition is different during the lunch breaks). For each state's static network, we compute the number of links that take place inside a group, l_{intra}, and those connecting different groups, l_{inter}. The average density of intra-groups links is then given by

$$p_{intra} = \frac{l_{intra}}{\sum_g \frac{n_g(n_g-1)}{2}}, \qquad (5)$$

with n_g the the number of nodes of group g, and for the inter-link density we obtain

$$p_{inter} = \frac{l_{inter}}{\sum_g \frac{n_g(N-n_g)}{2}}, \qquad (6)$$

where the denominators correspond to the maximal possible values of l_{intra} and l_{inter}, respectively. We define the parameter χ as p_{inter}/p_{intra} so it quantifies the probability of having links between groups with respect to inside a group.

Generation of the first d snapshots of the surrogate
The first snapshot of the surrogate cannot be based on previous interactions so we build it using the configuration model[4], which allows us to generate a static network with a given degree distribution, and we choose the degree distribution of the first snapshot of the original network. To generate the second snapshot we can then rely on the interactions of the first snapshot, so we use the same method described above but using temporal neighborhoods of length 2 timesteps instead of $d+1$. We repeat the procedure d times, until we have the d snapshots that we need to initialize the process.

Temporal states clustering
In the section "Decomposition in temporal states", we have briefly described how to decompose the temporal network into temporal states, each composed of a set of (non-necessarily contiguous) temporal snapshots. More in details, once the distances have been computed between all pairs of snapshots, we can cluster them into C clusters by a hierarchical clustering with a

bottom up approach: we start by considering each snapshot as a different cluster, and we proceed by iteratively merging the clusters. We do this for $C \in [3, T]$ for the Laplacian distance and for $C \in [4, T]$ for the activity distance, thus avoiding all the solutions with only a few clusters that would not provide sufficient state variability. We then chose among all these possible partitions the one that maximizes the Dunn's index[54], defined as

$$\frac{\min_{1 \leq c \neq c' \leq C} \min_{i \in c\text{th state}, j \in c'\text{th state}} d(i, j)}{\max_{1 \leq c'' \leq C} \max_{i', j' \in c''\text{th state}} d(i', j')}, \qquad (7)$$

where $d(i, j)$ is either $d^{\mathcal{L}}$ or d^{e}. The numerator is the smallest distance between two states among all pairs of states and the denominator is the largest distance inside a state among all states.

Absent nodes
The population of a dataset of social interactions can be more or less stable: for some datasets the same set of individuals is present during the entire time span, for other datasets individuals disappear and new individuals appear making the population vary with time.

The method that we have described to generate surrogate networks does not include such a variability of the population, and in principle all nodes can be active at any time. For this reason we include a possible variation, to be applied to datasets where the population change is important. In fact, if in the original network we observe that some nodes do not participate to the dataset during specific time intervals, we can reproduce this feature in the surrogate networks too. This is implemented by imposing that during the same time intervals these nodes, instead of sampling their ego-subgraph from the probability distribution, always choose the empty ego-subgraph, with no interactions. These nodes will not participate to interactions, hence appearing as absent, in specific intervals of the generated network, reproducing the original network population variability.

SIR model: R_0 and R_∞
In the SIR model we compute two observables that are commonly used to evaluate the disease spreading dynamics. The first is the basic reproductive number R_0, that is numerically computed by counting the number of direct contagions due to the first infected node, i.e., the number of nodes that are infected by this seed node and not by other nodes. The second observable is the final number of recovered, R_∞, that is counted considering only the simulations where at least one contagion event takes place (we exclude all the realizations where $R_\infty = 1$, which corresponds to runs in which the seed recovered before infecting anybody else).

Deffuant model and Naming Game on temporal networks
The Deffuant model of opinion dynamics[49] and the Naming Game[51,52] have been initially formulated for static networks, and can be generalized to temporal ones in several ways[50,65]. In our implementation of both processes, at each snapshot each couple of connected nodes (taken in random order) is considered once and their opinions x or their name inventory are immediately updated (or not). This means that each node at a generic snapshot can change opinion or change its name inventory as many times as its degree.

For instance, for the Deffuant model, if a node i is connected to node j and l at time t, the interaction with j will change the opinion of i from x_i to x_i' and the interaction with l will take place with the updated opinion x_i', that will again change to x_i'' because of l. The final opinion of i at time t will be x_i''.

The same connections are similarly treated in the Naming Game: first the roles of speaker and hearer are randomly assigned to i and j. Then their name inventory are updated due to their interaction, and the new name inventory of i will be used in the name exchange with l.

The order in which connections are considered at each time step plays a role in the final result. For the sake of generality we choose a random order and repeat the process 100 times to explore a large set of possible evolution patterns.

References
1. Barrat, A., Barthelemy, M. & Vespignani, A. *Dynamical Processes on Complex Networks* (Cambridge University Press, 2008).
2. Pósfai, M. & Barabási, A.-L. *Network science* (Cambridge University Press, 2016).
3. Latora, V., Nicosia, V. & Russo, G. *Complex networks: principles, methods and applications* (Cambridge University Press, 2017).
4. Newman, M. *Networks* (Oxford university press, 2018).
5. Holme, P. & Saramäki, J. Temporal networks. *Phys. Rep.* **519**, 97–125 (2012).
6. Masuda, N. & Lambiotte, R. *A guide to temporal networks* (World Scientific, 2016).
7. Castellano, C., Fortunato, S. & Loreto, V. Statistical physics of social dynamics. *Rev. Mod. Phys.* **81**, 591–646 (2009).
8. Pastor-Satorras, R., Castellano, C., Van Mieghem, P. & Vespignani, A. Epidemic processes in complex networks. *Rev. Mod. Phys.* **87**, 925 (2015).
9. Eames, K., Bansal, S., Frost, S. & Riley, S. Six challenges in measuring contact networks for use in modelling. *Epidemics* **10**, 72–77 (2015).
10. Génois, M., Vestergaard, C. L., Cattuto, C. & Barrat, A. Compensating for population sampling in simulations of epidemic spread on temporal contact networks. *Nat. Commun.* **6**, 8860 (2015).
11. Stehlé, J. Simulation of an SEIR infectious disease model on the dynamic contact network of conference attendees. *BMC Med.* **9**, 87 (2011).
12. Calmon, L., Colosi, E., Bassignana, G., Barrat, A. & Colizza, V. Preserving friendships in school contacts: an algorithm to construct synthetic temporal networks for epidemic modelling. *PLoS Comput. Biol.* **20**, e1012661 (2024).
13. Braha, D. & Bar-Yam, Y. From centrality to temporary fame: dynamic centrality in complex networks. *Complexity* **12**, 59–63 (2006).
14. Isella, L. What's in a crowd? analysis of face-to-face behavioral networks. *J. Ther. Biol.* **271**, 166–180 (2011).
15. Masuda, N. & Holme, P. Detecting sequences of system states in temporal networks. *Sci. Rep.* **9**, 795 (2019).
16. Galimberti, E. et al. Span-core decomposition for temporal networks: Algorithms and applications. *ACM Transactions on Knowledge Discovery from Data (TKDD)* **15**, 1–44 (2020).
17. Pedreschi, N., Battaglia, D. & Barrat, A. The temporal rich club phenomenon. *Nat. Phys.* **18**, 931–938 (2022).
18. Hartle, H. & Masuda, N. Autocorrelation properties of temporal networks governed by dynamic node variables. *Phys. Rev. Res.* **7**, 013083 (2025).
19. Karsai, M., Kaski, K., Barabási, A.-L. & Kertész, J. Universal features of correlated bursty behaviour. *Sci. Rep.* **2**, 397 (2012).
20. Perra, N., Gonçalves, B., Pastor-Satorras, R. & Vespignani, A. Activity driven modeling of time varying networks. *Sci. Rep.* **2**, 469 (2012).
21. Laurent, G., Saramäki, J. & Karsai, M. From calls to communities: a model for time-varying social networks. *Eur. Phys. J. B* **88**, 1–10 (2015).

22. Starnini, M. & Pastor-Satorras,R. Topological properties of a time-integrated activity-driven network. *Phys. Rev. E-Stat. Nonlinear Soft Matter Phys.* **87**, 062807 (2013).

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

24. Le Bail, D., Génois, M. & Barrat, A. Modeling framework unifying contact and social networks. *Phys. Rev. E* **107**, 024301 (2023).

25. Stehlé, J., Barrat, A. & Bianconi, G. Dynamical and bursty interactions in social networks. *Phys. Rev. E Stat. Nonlinear Soft Matter Phys.* **81**, 035101 (2010).

26. Zhao, K., Stehlé, J., Bianconi, G. & Barrat, A. Social network dynamics of face-to-face interactions. *Phys. Rev. E-Stat. Nonlinear Soft Matter Phys.* **83**, 056109 (2011).

27. Moinet, A., Starnini, M. & Pastor-Satorras, R. Burstiness and aging in social temporal networks. *Phys. Rev. Lett.* **114**, 108701 (2015).

28. Moinet, A., Starnini, M. & Pastor-Satorras, R. Aging and percolation dynamics in a non-poissonian temporal network model. *Phys. Rev. E* **94**, 022316 (2016).

29. Kim, H., Ha, M. & Jeong, H. Dynamic topologies of activity-driven temporal networks with memory. *Phys. Rev. E* **97**, 062148 (2018).

30. Williams, O. E., Lacasa, L., Millán, A. P. & Latora, V. The shape of memory in temporal networks. *Nat. Commun.* **13**, 499 (2022).

31. Rocha, L. E. & Blondel,V. D. Bursts of vertex activation and epidemics in evolving networks. *PLoS Comput. Biol.* **9**, e1002974 (2013).

32. Vestergaard, C. L., Génois, M. & Barrat, A. How memory generates heterogeneous dynamics in temporal networks. *Phys. Rev. E* **90**, 042805 (2014).

33. Jo, H.-H., Pan, R. K. & Kaski, K. Emergence of bursts and communities in evolving weighted networks. *PloS ONE* **6**, e22687 (2011).

34. Purohit, S., Holder, L. B. & Chin, G. Temporal graph generation based on a distribution of temporal motifs. In *Proceedings of the 14th International Workshop on Mining and Learning with Graphs*, 7 (2018).

35. Zhou, D., Zheng, L., Han, J. & He, J. A data-driven graph generative model for temporal interaction networks. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 401–411 (2020).

36. Presigny, C., Holme, P. & Barrat, A. Building surrogate temporal network data from observed backbones. *Phys. Rev. E* **103**, 052304 (2021).

37. Zeno, G., La Fond, T. & Neville, J. Dymond: dynamic motif-nodes network generative model. In *Proc. Web Conference 2021*, 718–729 (2021).

38. Duval, A., Leclerc, Q. J., Guillemot, D., Temime, L. & Opatowski, L. An algorithm to build synthetic temporal contact networks based on close-proximity interactions data. *PLOS Comput. Biol.* **20**, e1012227 (2024).

39. Gelardi, V., Le Bail, D., Barrat, A. & Claidiere, N. From temporal network data to the dynamics of social relationships. *Proc. R. Soc. B* **288**, 20211164 (2021).

40. Longa, A., Cencetti, G., Lehmann, S., Passerini, A. & Lepri, B. Generating fine-grained surrogate temporal networks. *Commun. Phys.* **7**, 22 (2024).

41. Fortunato, S. Community detection in graphs. *Phys. Rep.* **486**, 75–174 (2010).

42. Barrat, A. & Cattuto, C. *Face-to-Face Interactions* (Springer International Publishing, 2015).

43. Stehlé, J. High-resolution measurements of face-to-face contact patterns in a primary school. *PLOS ONE* **6**, e23176 (2011).

44. Fournet, J. & Barrat,A. Contact patterns among high school students. *PloS ONE* **9**, e107878 (2014).

45. Mastrandrea, R., Fournet, J. & Barrat, A. Contact patterns in a high school: a comparison between data collected using wearable sensors, contact diaries and friendship surveys. *PloS ONE* **10**, e0136497 (2015).

46. Génois, M. & Barrat,A. Can co-location be used as a proxy for face-to-face contacts? *EPJ Data Sci.* **7**, 11 (2018).

47. Toth, D. J. The role of heterogeneity in contact timing and duration in network models of influenza spread in schools. *J. R. Soc. Interface* **12**, 20150279 (2015).

48. Sapiezynski, P., Stopczynski, A., Lassen, D. D. & Lehmann, S. Interaction data from the copenhagen networks study. *Sci. Data* **6**, 1–10 (2019).

49. Deffuant, G., Neau, D., Amblard, F. & Weisbuch, G. Mixing beliefs among interacting agents. *Adv. Complex Syst.* **3**, 87–98 (2000).

50. Zarei, F., Gandica, Y. & Rocha, L. E. Bursts of communication increase opinion diversity in the temporal deffuant model. *Sci. Rep.* **14**, 2222 (2024).

51. Baronchelli, A., Dall'Asta, L., Barrat, A. & Loreto, V. Nonequilibrium phase transition in negotiation dynamics. *Phys. Rev. E-Stat., Nonlinear, Soft Matter Phys.* **76**, 051102 (2007).

52. Baronchelli, A. A gentle introduction to the minimal naming game. *Belgian J. Linguist.* **30**, 171–192 (2016).

53. Longa, A., Cencetti, G., Lepri, B. & Passerini, A. An efficient procedure for mining egocentric temporal motifs. *Data Min. Knowl. Discov.* **36**, 355–378 (2022).

54. Dunn, J. C. A fuzzy relative of the isodata process and its use in detecting compact well-separated clusters. *Journal of Cybernetics* **3**, 32–57 (1973).

55. Blondel, V. D., Guillaume, J.-L., Lambiotte, R. & Lefebvre, E. Fast unfolding of communities in large networks. *J. Stat. Mech. Theory Exp.* **2008**, P10008 (2008).

56. Kobler, J., Schöning, U. & Torán, J. *The graph isomorphism problem: its structural complexity* (Springer Science & Business Media, 2012).

57. Newman, M. E. J. Mixing patterns in networks. *Phys. Rev. E* **67**, 026126 (2003).

58. Battiston, F. Networks beyond pairwise interactions: Structure and dynamics. *Phys. Rep.* **874**, 1–92 (2020).

59. Iacopini, I., Petri, G., Barrat, A. & Latora, V. Simplicial models of social contagion. *Nat. Commun.* **10**, 2485 (2019).

60. Goh, K.-I. & Barabási,A.-L. Burstiness and memory in complex systems. *Europhys. Lett.* **81**, 48002 (2008).

61. Karsai, M. et al. *Bursty human dynamics* (Springer, 2018).

62. Ciaperoni, M. Relevance of temporal cores for epidemic spread in temporal networks. *Sci. Rep.* **10**, 12529 (2020).

63. Valdano, E. Predicting epidemic risk from past temporal contact data. *PLOS Comput. Biol.* **11**, 1–19 (2015).

64. Loedy, N., Wallinga, J., Hens, N. & Torneri, A. Repetition in social contacts: implications in modelling the transmission of respiratory infectious diseases in pre-pandemic and pandemic settings. *Proc. R. Soc. B: Biol. Sci.* **291**, 20241296 (2024).

65. Maity, S. K., Manoj, T. V. & Mukherjee, A. Opinion formation in time-varying social networks: the case of the naming game. *Phys. Rev. E-Stat. Nonlinear Soft Matter Phys.* **86**, 036110 (2012).

66. Lance, G. N. & Williams,W. T. Computer programs for hierarchical polythetic classification ("similarity analyses"). *Comput. J.* **9**, 60–64 (1966).

Acknowledgements

Author contributions

G.C. contributed in conceptualization, formal analysis, investigation, and writing. A.B. contributed in supervision, investigation, and writing.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s42005-025-02075-4.

Correspondence and requests for materials should be addressed to Giulia Cencetti.

Peer review information *Communications Physics* thanks the anonymous reviewers for their contribution to the peer review of this work. A peer review file is available.

Reprints and permissions information is available at http://www.nature.com/reprints

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