Building surrogate temporal network data from observed backbones

Charley Presigny,[1] Petter Holme,[2] and Alain Barrat[1,2,*]

[1]*Aix Marseille Univ, Université de Toulon, CNRS, CPT, Turing Center for Living Systems, 13288 Marseille, France*

[2]*Tokyo Tech World Research Hub Initiative (WRHI), Tokyo Institute of Technology, Yokohama 226-8503, Japan*

(Received 23 December 2020; accepted 22 April 2021; published 6 May 2021)

Many systems of socioeconomic interests find a convenient representation in the form of temporal networks, i.e., sets of nodes and interactions occurring at specified times. In the corresponding data sets, however, crucial elements coexist with nonessential ones and noise. Several methods have thus been proposed to extract a "network backbone," i.e., the set of most important links in a network data set. The outcome of such methods can be seen as compressed versions of the original data. However, the question of how to practically use such reduced views of the data has not been tackled: for instance, using them directly in numerical simulations of processes on networks might lead to important biases. Overall, such reduced views of the data might not be actionable without an adequate decompression method. Here, we address this issue by putting forward and exploring several systematic procedures to build surrogate data from various kinds of temporal network backbones. In particular, we explore how much information about the original data needs to be retained alongside the backbone so that the surrogate data can be used in data-driven numerical simulations of spreading processes on a wide range of spreading parameters. We illustrate our results using empirical temporal networks with a broad variety of structures and properties. Our results give hints on how to best summarize complex data sets so that they remain actionable. Moreover, they show how ensembles of surrogate data with similar properties can be obtained from an original single data set, without any modeling assumptions.

DOI: 10.1103/PhysRevE.103.052304

I. INTRODUCTION

Many data sets coming from the world around us—transportation systems, human proximity, interactions on social media, etc.—take the form of networks [1,2]. By network, we mean any system that could be modeled meaningfully as a graph $G = (V, E)$ of nodes V and edges E (pairs of nodes). Often one also knows when nodes are in contact, which calls for a temporal network approach [3–5] where the set of edges is replaced by a set of contacts—triples giving pairs of nodes and the time of interaction. One of the main objectives in the study of networks is to describe a graph G succinctly, and from this description be able to regenerate other graphs with statistical properties as close as possible to those of G. There are many approaches to this endeavor, depending on how brief we want the description to be. Indeed, there is a continuum, from few-parameter models where the reconstructed networks necessarily will be different from the observed system to link prediction where the objective is to add just one link to an otherwise completely specified graph. This paper concerns the intermediate regime of this spectrum that Ref. [6] calls *graph summarization techniques*.

In graph summarization, one allows a description to be long enough to contain a non-negligible part of the original network. Thus, one could describe the network by its most important links and edges and complement this description with a simple model for how to regenerate an entire network [7,8] (see Fig. 1). There are many ultimate uses of such a method [6]: It could save space when storing large data. It could be statistically more relevant to generate uncertain links by a model. It could speed up algorithms.

Graph summarization is intimately linked to identifying the most important subgraph of G, often called *backbone extraction* [9–13]. In the present work, we seek to extend backbone extraction methods for temporal networks to a full graph summarization technique. We will refer to this extension as the creation of *surrogate temporal networks*. We will primarily try to construct surrogate networks that predict the same epidemic outbreaks as the original data. Indeed, networks are the substrate of many dynamical processes, among which epidemic processes are a prominent example [2], and an important use of network data sets is indeed to provide support for data-driven simulations of these processes. Surrogate networks should thus in particular provide data that can be used in lieu of the original data in such simulations.

Several methods have been put forward to extract network backbones. For static weighted networks, the simplest way of filtering edges is to remove all the edges with weight below a given threshold value. More principled procedures use statistical tests based on null models to compare the weights of the edges with the ones that would be generated at random by a certain null model [9–13]. One then fixes a desired significance level and selects only those edges whose weight cannot be explained by the null model at the chosen significance level. These significant edges form the backbone of the network. In the case of temporal networks, a simple approach for the extraction of backbones is to aggregate the

[*] alain.barrat@cpt.univ-mrs.fr

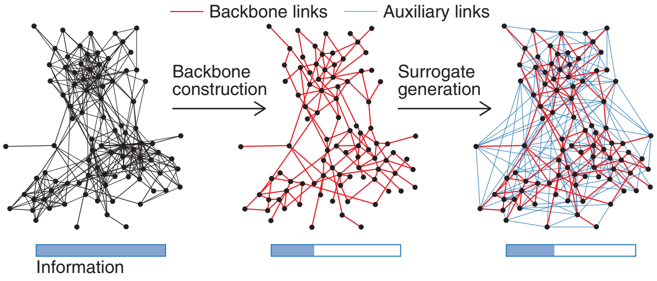

FIG. 1. Illustration of backbones and surrogates. The backbone construction identifies the most important links, and thereby compresses the original data. The surrogate generation model adds auxiliary links, extracted at random using specific procedures, to create a network of the same size as the original. The more links (information) retained in the backbone construction, the more similar is the surrogate data to the original.

data into a weighted static network. The weight of a link in this network is the number, or total duration, of the interactions between the involved nodes. To avoid neglecting potentially critical temporal features, it is, however, necessary to define an adequate temporal null model. Such a procedure makes it possible to extract a backbone of significant ties, i.e., of meaningful sequences of temporal contacts between nodes, possibly taking into account the temporal evolution of the nodes' properties [14–16].

Typical backbone-method studies validate the procedures to extract backbones from static and temporal networks on synthetic benchmark tests and various empirical data sets. One explores the main properties of the resulting backbones, and compares these to known properties to understand by which network features they are influenced. However, there is most often no explicit interpretation of the "importance" of the links (except for the simple weight-thresholding procedure). Most importantly, it is unknown whether the information contained in the extracted backbone is enough to correctly summarize the original data and to be actionable, i.e., whether a user with access to the backbone but not to the original data can use it in data-driven applications such as simulations of dynamical processes.

In this paper, we explore this issue for backbone methods in temporal networks. Given a backbone representing only a fraction of the original data, we put forward and explore several systematic procedures to reconstruct surrogate actionable data by adding auxiliary links to the backbone (see Fig. 1) [7,8]. These auxiliary links are extracted at random with a procedure depending on how the backbone was created. We compare several such procedures applied to backbones obtained through a simple thresholding procedure (serving as baseline) and the significant tie (ST) filter for temporal networks [14]. We also propose a new version of this filter that considers the data's potential group structure. In each case, we explore how much information about the original data needs to be kept alongside the backbone (e.g., some statistical properties concerning the links that have been filtered out). To show our results' generality, we study temporal-network data with a broad range of topological and temporal structures.

II. DATA AND GENERAL METHODOLOGY

We consider data sets describing contacts between individuals with temporal resolution, collected by the SocioPatterns collaboration [17] in different settings: a workplace (office building, InVS15) [18], a high school (Thiers13) [19], a primary school (LyonSchool) [20], and a scientific conference (SFHH) [21]. These data describe close face-to-face proximity of individuals equipped with wearable sensors, with a temporal resolution of 20 s. To limit the effect of noise, the data are, moreover, often aggregated over a coarser resolution of Δ minutes (e.g., in Ref. [14] backbones are considered for Δ ranging from 3 to 15 min). Here we will use $\Delta = 3$ min, but we have obtained similar results for other temporal resolutions. Such data are conveniently represented as temporal networks in which nodes represent individuals. These networks are in discrete time, i.e., composed by T successive snapshots at times t_0, $t_0 + \Delta, \ldots$, where t_0 is the initial time of the data set. A temporal edge between two nodes i and j at time $t = t_0 + n\Delta$ represents the fact that the corresponding nodes have been in contact during the time interval $[t, t + \Delta]$. We also define a "contact" between i and j as an uninterrupted series of time stamps in which there is a temporal edge between them. The duration of the contact is the length of this series. In each case, we also define the aggregated network as the static weighted network in which a link between two nodes denotes that these two nodes have been in contact at least once, and the weight of the link is given by the number of temporal edges between these nodes. Table I gives the main characteristics of each data set.

These data sets were collected in very different contexts, so that the resulting structural and temporal properties of the contact network differ strongly. School and high school populations are divided into classes of similar sizes, with a strong community structure and interactions between classes only during the breaks (occurring with similar patterns in different days) [19,20]. In the office building, individuals are divided into departments of unequal sizes, and interactions are not limited by strict schedules [21]. In the conference, a homogeneous aggregated contact network is observed [22].

For each data set, we first extract their backbones according either to a simple thresholding procedure or using the significant tie filter [14] (see below and Methods). Each backbone contains only a tunable fraction f of the original ties (we will use $f = 40\%$, 10%, and 5%). In addition to the list of backbone ties (and possibly the corresponding lists of temporal edges), we assume that some additional statistics of the original data sets are conserved, such as the total number of temporal edges, and the distributions of contact and intercontact durations (or simply the parameters of their fit to simple functional forms such as power laws [7,23,24]). Whenever the data present a group structure, the corresponding metadata can also be conserved alongside the backbones.

We then consider several methods to reconstruct surrogate data from the backbones. Each method consists in adding temporal edges to the backbone in a way tailored to reproduce several statistical features of the original data (see below and Methods). For the resulting surrogate data, we investigate whether they are suitable to feed numerical simulations of dynamical processes, i.e., whether the outcome of dynamical

TABLE I. The basic properties of the different data sets. N is the number of participants, "Duration" the total duration of the data collection, N_g the number of groups in the population, E the number of ties (i.e., links in the aggregated network), T the number of time stamps (once nights and weekends, with no activity, have been removed), and E_T the number of temporal edges. Here the temporal resolution is $\Delta = 3$ min.

Data set	Location	Year	N	Duration	N_g	E	T	E_T	Ref.
InVS15	Office building	2015	217	2 weeks	12	4274	2307	28 950	[21]
LyonSchool	Primary school	2009	242	2 days	10	8317	345	64 419	[20]
SFHH	Conference	2009	403	2 days	None	9565	421	73 620	[21]
Thiers13	High school	2013	326	1 week	9	5818	811	59 372	[19]

processes simulated on top of the surrogate data is close to the one obtained when using the original data. Specifically, we focus on the paradigmatic susceptible-infectious-recovered (SIR) model of epidemic propagation. In this model, a susceptible (S) node becomes infectious (I) at rate β when in contact with an infectious node. Infectious nodes recover spontaneously at rate ν and enter an immune recovered (R) state. We quantify the outcome of these processes, i.e., the epidemic risk, by two quantities: (i) the basic reproductive number R_0 (the average number of secondary infections by the source) and (ii) the average final size Ω of the spread, i.e., the fraction of nodes that have been in the infectious state at any time, and we explore a wide range of parameter values (see Methods for details on numerical simulations and measures).

In the following, we will show in the main text the results for the Thiers13 data set. As we indeed observe a robust phenomenology across data sets, the results for the other data sets are shown in the Supplemental Material [25].

III. BACKBONES

To extract a backbone of a given size from a temporal network data set, we consider the ST filter [14]. In this method, the actual number of temporal edges between two nodes is compared to the one of a temporal null model. The significant ties at significance level α are the ones such that their number of temporal edges cannot be explained by the null model at significance level α. Specifically, the null model is defined as follows: an "activity level" a_i is associated to each node i, and two nodes i and j have a temporal edge at each time with probability $a_i a_j$. The activity levels of the nodes are obtained from the data by maximum likelihood estimation (see Methods and Ref. [14]). Tuning α makes it possible to select backbones representing a specific fraction f of the ties of the original data.

Moreover, we extend the ST filter to take into account the group structure of several data sets. The resulting GST filter is obtained by modifying the temporal null model as follows: the probability of a temporal edge between i and j is equal to $a_i a_j$ if i and j belong to the same group, and to $pa_i a_j$ if they belong to different groups. The node activities and parameter p are obtained by maximum likelihood estimation as for the ST filter (see Methods). Note that $p < 1$ corresponds to cohesive group structures, while $p > 1$ would be obtained for disassortative structures. It would also be possible to use several values of p depending on the respective groups of i and j, but we consider here for simplicity only one parameter.

In addition, we consider as baseline the simplest method to extract ties that can be interpreted as the most important in a network: We order the ties according to their weight in the aggregated network, as given by their number of temporal edges (in the context of contact networks, this corresponds to the total duration of the contacts between the two nodes forming the tie). The "threshold" backbone (TB) of the original data is then given by the fraction f of ties with the largest weights.

We report in Table II, for backbones formed of a fraction $f = 40\%$, 10%, and 5% of the original network, the corresponding number of temporal edges for each backbone extraction method. See the Supplemental Material [25] for more information about how backbones compare to the original data. As already discussed in Ref. [14], the ST backbone ties tend to have large weights, with distributions clearly shifted to large values with respect to the original data. However, while this happens by definition in the threshold backbone, the distribution of weights in the ST backbone is smooth and does not have a sharp cutoff at a minimal value. Moreover, when the group structure is included (GST backbone), the distribution of weights becomes notably broader. This is due to intergroup ties that tend to have lower weights [7]: these ties appear as significant only when we take into account, through the adequate null model (i.e., through the use of the parameter p), that pairs of individuals belonging to different groups have an *a priori* tendency to form less temporal edges than individuals of the same group. In fact, the ST filter tends to filter out most ties joining nodes of different groups [14]; the GST filter instead keeps ties both within and between groups. We also note that both the clustering coefficient and the modularity of the partition in groups, when measured in the backbones, can strongly deviate from the values in the original data (see tables in the Supplemental Material [25]). On the other hand, the distributions of contact and intercontact durations are close to the ones observed in the original data (see the Supplemental Material [25]).

IV. CONSTRUCTING SURROGATE DATA

Backbones are by definition composed of a much smaller number of temporal edges and ties than the original data. As discussed above, their statistical properties are not identical to the ones of the data. It is therefore expected that numerical simulations of spreading processes on top of a backbone largely underestimate their outcome. We illustrate this in Fig. 2 and in the Supplemental Material [25]. Note that the underestimation is not as strong as the one that would be obtained by a random sampling of the events, as the backbone ties tend to have large weights.

TABLE II. Number of temporal edges E_{bT} of the various backbones, for various values of the fraction f of ties forming the backbone.

Data set		Threshold			ST			GST	
f	40%	10%	5%	40%	10%	5%	40%	10%	5%
InVS15	25 417	16 541	12 014	24 738	16 166	9 084	20 890	13 581	9 522
LyonSchool	56 807	33 205	22 912	55 773	32 346	18 746	41 032	24 732	16 431
SFHH	18 802	12 650	10 253	17 257	11 950	9 763			
Thiers13	53 992	39 171	30 239	52 975	30 981	12 678	43 834	34 613	22 572

We therefore put forward several methods to construct surrogate data that are statistically more similar to the original data and, most importantly, yield more accurate estimations of processes' outcomes. Starting from a backbone composed of E_b ties and E_{bT} temporal edges, we want to recreate a temporal network with approximately E ties and E_T temporal edges. To this aim, we need to use complementary information, in addition to the list of temporal edges composing the backbone. For instance, it is quite clear that we cannot guess from the backbone itself the correct numbers of ties and temporal edges to be added. Thus, this additional information should be kept alongside the backbone to make it a usable summary of the data. Here we consider several procedures, highlighting in each case the necessary type and amount of information. Note that the resulting list of procedures does not pretend to be exhaustive but addresses a wide range of possibilities in terms of available information. Each procedure can be separated into two steps: (i) choosing ties (not included in the backbone) that interact in the surrogate data, and (ii) building timelines of interactions on the chosen ties. Procedure (ii) might also need to be performed on the backbone ties if the temporal information of the backbone ties is not available.

For step (i), we consider three distinct methods for backbones extracted using the ST or GST method.

(G)ST-OA. Here "OA" stands for "original activity." We assume that the parameters of the null model used to extract the backbones are available, namely, the original node activities $\{a_i, i = 1, \ldots, N\}$ (and the parameter p for the GST). Moreover, we assume that E_T is known and, for the GST, that the number of temporal edges between groups and within groups, $E_{T,\text{inter}}$ and $E_{T,\text{intra}}$, are also known, as well as the group to which each node belongs.

In this procedure, for each pair of nodes (i, j) not in the backbone, we add a temporal edge between i and j at each time stamp with probability $\alpha a_i a_j$, calibrating α so that the obtained total number of temporal edges is close to E_T (see Methods).

For the GST, we use at each time the probabilities $\alpha_{\text{intra}} a_i a_j$ if i and j are in the same group and $\alpha_{\text{inter}} p a_i a_j$ if they are not, calibrating α_{inter} and α_{intra} to get approximately the correct number of temporal edges both at the intergroup and intragroup levels.

(G)ST-RA. Here "RA" stands for "recomputed activity." If the parameters of the null model (i.e., the activities $\{a_i\}$ of the nodes) are not known, we use the fact that applying the MLE equations to the backbone itself yields activity parameters correlated to the original ones (see Table III). We thus compute the activity $\tilde{a}_i$ of each node i (and the parameter $\tilde{p}$ if the group

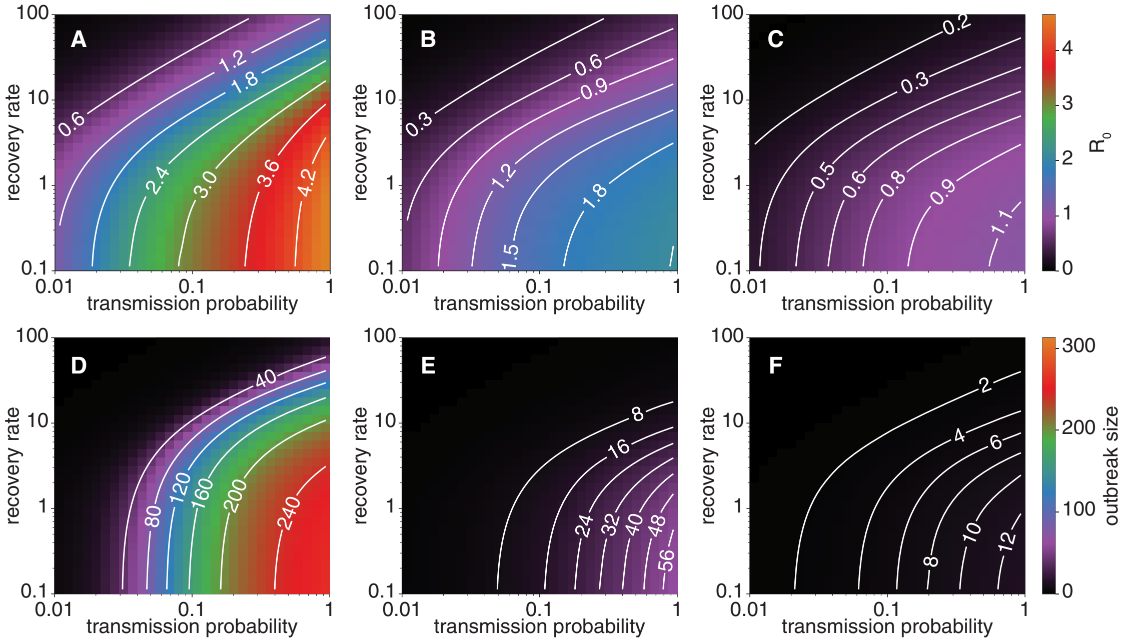

FIG. 2. R_0 (top row) and Ω (bottom row) values obtained from the simulations on the original data and on the backbones, as a function of the SIR parameters, for the Thiers13 data set. Left: original data. Middle: GST backbone with $f = 10\%$. Right: GST backbone with $f = 5\%$.

TABLE III. Correlation between original activities and activities recomputed using the backbone ties.

Data set		ST			GST	
f	40%	10%	5%	40%	10%	5%
InVS15	0.99	0.92	0.62	0.97	0.87	0.75
LyonSchool	0.99	0.90	0.60	0.96	0.80	0.64
SFHH	0.97	0.93	0.90			
Thiers13	0.99	0.73	0.25	0.97	0.92	0.68

structure is known) in the backbone; we then add at each time a temporal edge between i and j with probability $\alpha\tilde{a}_i\tilde{a}_j$, calibrating α to get approximately the correct number of temporal edges (we assume as previously that E_T is known). For the GST case, we use probabilities $\alpha_{\text{intra}}\tilde{a}_i\tilde{a}_j$ and $\alpha_{\text{inter}}\tilde{p}\tilde{a}_i\tilde{a}_j$ and calibrate $\alpha_{\text{inter/intra}}$ as in the previous method, assuming $E_{T,\text{inter}}$ and $E_{T,\text{intra}}$ are known.

(G)ST-RT. Here "RT" stands for "random tie." We, moreover, consider a baseline in which we add to the backbone the correct number of ties at random (i.e., $E - E_b$), with weights drawn from the list of weights of the nonbackbone ties. Note that here we do not consider simply adding the correct number of temporal edges at random between nodes, because that would result in a very large number of ties with only one or few temporal edges, a structure very different from the original data. We thus assume that the number of ties in the original data E is known (E_{inter} and E_{intra} if there are groups), in addition to the original number of temporal edges. Moreover, as the distribution of the backbone weights is very different from the original data (see figures in the Supplemental Material [25] and Ref. [14]), we do not have a simple functional form for the weights of the nonbackbone ties. We thus assume that the list of weights of the nonsignificant ties has been kept.

Finally, for the backbones consisting of the ties with the largest weights (TB), as there is no underlying null model, we only consider the baseline reconstruction method which we denote by (G)TB-RT: we proceed here exactly as for the (G)ST-RT procedure.

Once the ties and the number of temporal edges on each tie have been chosen by one of these procedures, we can create surrogate timelines [step (ii) of the procedure] in various ways. In each case, for each tie (i, j) with number of temporal edges n_{ij}, the aim is to choose n_{ij} time stamps out of the T possible ones.

Poisson. If no temporal information on the original data is available, the simplest procedure consists in choosing the time stamps of the temporal edges for each tie totally at random.

BTL-Poisson. If the actual timelines of the backbone ties are known, one can keep these actual timelines and choose at random the time stamps of temporal edges only for the surrogate ties.

Stats. Rather than the above approaches, we can instead assume that some information on the statistics of contact and intercontact durations are known, as these properties have been shown to be extremely robust [21,23] (see also the Supplemental Material [25]). They can, moreover, be approximately fitted to (truncated) power-law forms, meaning that the whole list of values is not needed, but only the parameters of the fit. We can then build a timeline of temporal edges for each tie using contact and intercontact durations generated randomly from these fitted distributions.

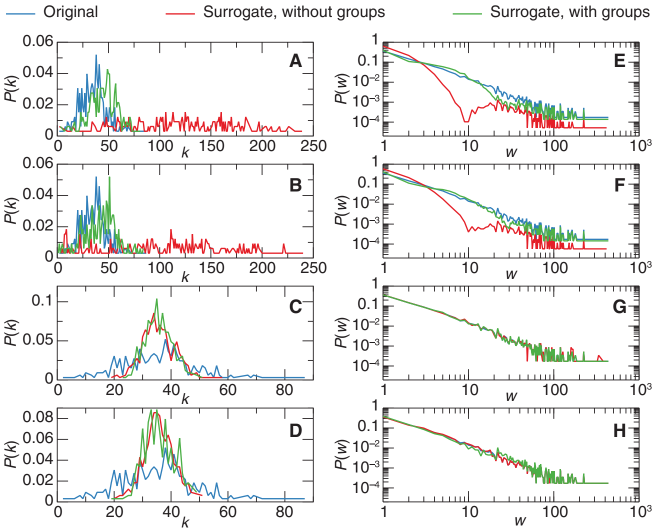

FIG. 3. Distributions of (aggregated) degrees (left columns) and weights (right column) in the surrogate data obtained by various methods. From top to bottom: (G)ST-OA, (G)ST-RA, (G)ST-RT, and (G)TB-RT. In each case, the blue line shows the distribution for the original data, and the red and green lines denote the surrogate built respectively without and with group information. Using group information yields distributions closer to the original ones. The surrogates were built from backbones with $f = 10\%$ of the ties of the original data.

BTL-Stats. If the actual timelines of the backbone ties are known, we keep these actual timelines, and proceed as in the Stats case for the surrogate ties only.

We note here that each step of the procedure is stochastic, with random choices of ties and temporal edges. Thus, repeating the same procedure multiple times yields an ensemble of surrogate temporal networks. In the Methods section, we provide a summary table of these procedures and the corresponding data used.

V. STRUCTURAL AND TEMPORAL STATISTICAL PROPERTIES OF THE SURROGATE DATA

Figure 3 shows distributions of degrees and weights for the aggregate networks resulting from surrogate data created by several methods for the Thiers13 data set and $f = 10\%$. Similar figures are shown in the Supplemental Material [25] for the other data sets as well as a table with the relative values of the clustering coefficient and of the modularity of the aggregated surrogate networks.

The surrogates based on adding ties according to the ST null model tend to overestimate the node degrees, with the whole distribution shifting to larger values than in the original data and becoming broader. This effect is very strong for the ST-OA and ST-RA, but taking into account groups (GST-OA and GST-RA) leads to much weaker deviations from the data. Using group data also leads to distributions of weights close to the original ones. At the same time, ST-OA and ST-RA have a substantial depletion of the distribution at intermediate weight

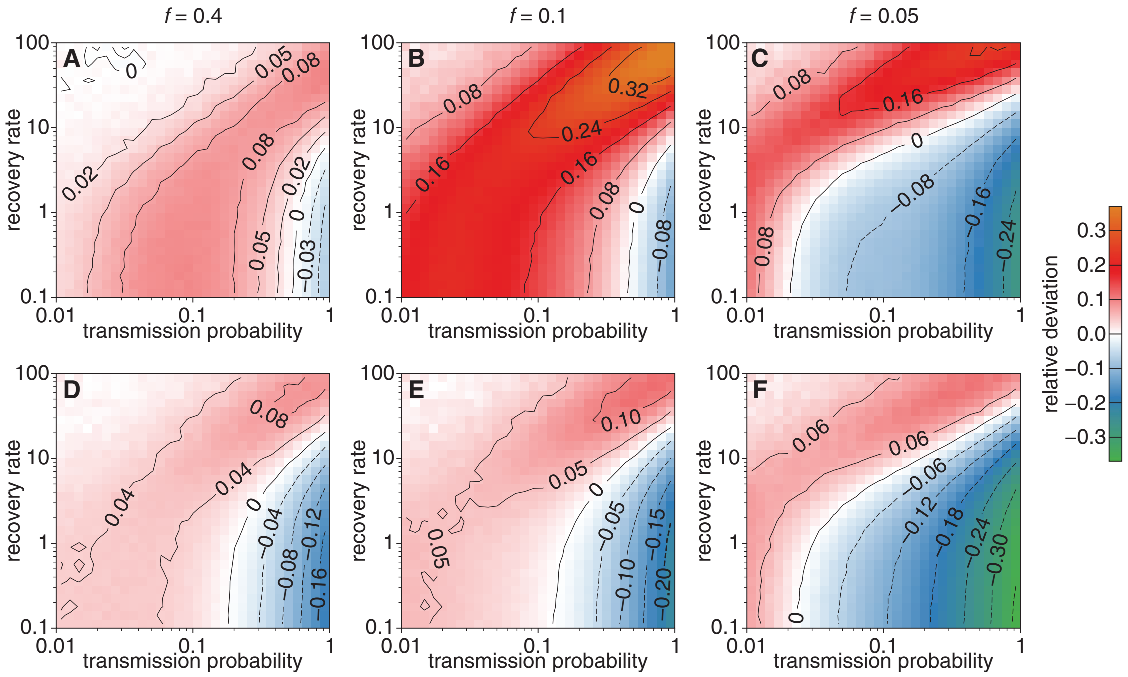

FIG. 4. Effect of taking group structure into account when building the surrogate data, for the Thiers13 data set. Each panel shows the relative difference in R_0 obtained from the simulations on surrogate data with respect to simulations on the original data. Top: ST-RA. Bottom: GST-RA. Left, $f = 40\%$; middle, $f = 10\%$; right, $f = 5\%$. In each case the backbone timelines are kept, and timelines respecting the statistics of contact and intercontact durations are built for the surrogate ties (BTL-Stats method).

values (the tail of the distribution being correctly represented as most ties with large weights belong to the ST backbone). Note that these distributions emerge from the surrogate's construction, as the initial distribution is not assumed to be known here.

For surrogates created using random ties, (G)ST-RT and (G)TB-RT, the average degree is well reproduced by design, as the information about the original data number of ties is assumed to be known. On the other hand, the distribution of degrees is much narrower than the original one. The distribution of weights is almost identical to the original data since the list of the actual weights of the nonsignificant ties is assumed to be known.

In terms of clustering and modularity, the procedures in which group information is known and used all lead to values that are close to the original ones, while ignoring group information can yield large discrepancies (see the Supplemental Material [25]).

Finally, the distributions of contact and intercontact durations depend only on the way in which the timelines of ties are built in the surrogate data: they are exponential for the Poisson procedure, and very close to the original data distributions for the Stats procedures (not shown).

VI. SIR PROCESSES ON SURROGATE DATA

We present our main results in Figs. 4 and 5. Each panel of the figures shows, as a function of the parameters β and ν, a color plot of the relative differences in the outcomes of SIR processes simulated either on surrogate data or on the original data. The outcome is measured by the basic reproductive number R_0, and we show similar results for the epidemic size Ω in the Supplemental Material [25]. The reason we do not reduce this objective measure to one number per data set is simply that any specific disease would correspond to one coordinate pair in the parameter space. Considering a summary statistics would hide the individual variability of different diseases even though it would simplify the analysis.

Let us first note a general pattern: R_0 tends to be underestimated, when using surrogate data, at large β and small ν, i.e., at very large R_0 and epidemic size. At large β and ν, on the other hand, the tendency is to overestimate the epidemic outcome. Finally, smaller deviations with respect to the simulations on the original data are observed in parameter regions where R_0 is close to 1, i.e., close to the epidemic threshold.

Let us now consider in more detail the results obtained with various types of surrogate data and the effect of the choices made in the reconstruction procedure.

Figures 4 (and further figures in the Supplemental Material [25]) highlight the impact of using information on the group structure of data. The surrogate data get more ties between groups when group information is not taken into account (see also the degree distribution in Fig. 3), leading to larger values of R_0 and Ω. As a result, the range of parameters in which R_0 is underestimated is slightly smaller. Still, on the other hand, both R_0 and Ω can be strongly overestimated in some parameter regions and in particular close to the epidemic threshold.

In the Supplemental Material [25], we furthermore examine the effect of different timeline reconstruction methods, at a fixed procedure for choosing the surrogate ties. As could be expected, better results are obtained when more statistical information about the actual data timelines is used. In particular, using Poisson timelines leads to stronger overestimations.

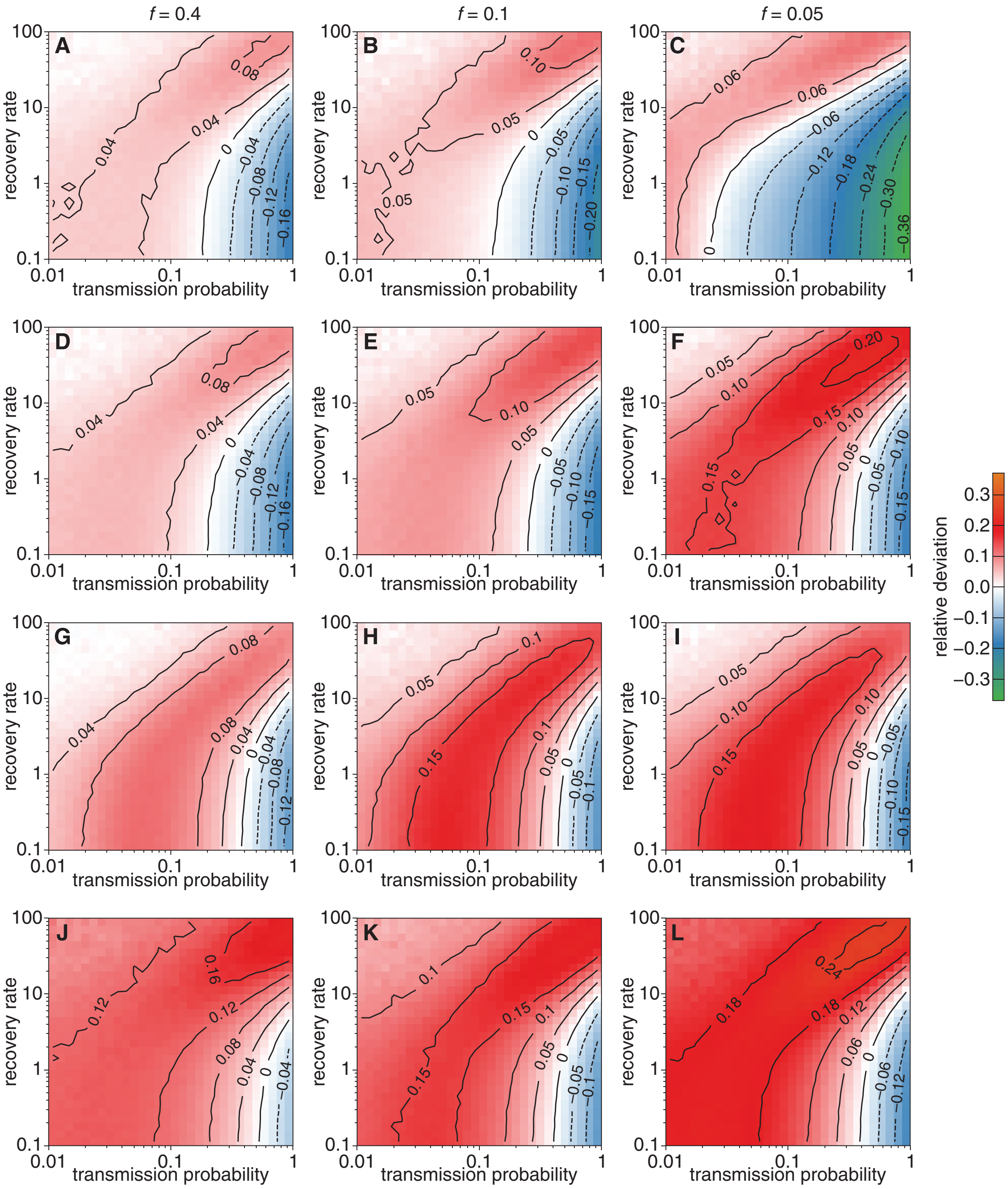

FIG. 5. Outcome of SIR processes on surrogate data obtained by various reconstruction methods, for the Thiers13 data set. Relative difference in the values of R_0 measured in simulations on the surrogate and on the original data. First row: GST-RA. Second row: GST-OA. Third row: GST-RT. Fourth row: GTB-RT. Left column, $f = 40\%$; middle, $f = 10\%$; right, $f = 5\%$. In each case the backbone timelines are kept, and timelines respecting the statistics of contact and intercontact durations are built for the surrogate ties (BTL-Stats method).

On the other hand, using timelines with random contact and intercontact durations reflecting the original data statistics leads to smaller deviations, and using these statistics to create surrogate timelines even on the backbone ties does not have a strong impact.

We thus consider all the surrogate reconstruction methods that take into account the group structure of the data and use the BTL-Stats method for the timelines. Figure 5 shows the results for R_0, while the results for Ω are shown in the Supplemental Material [25]. The main result underlined by the panels is that all methods give rather good results. The deviations with respect to the original data naturally tend to increase as f decreases, but wide ranges of parameters with small variations are still observed even at $f = 5\%$. We also see that recomputing the activities leads to worse underestimations than if the original activities are known. Despite being based on the simple procedure of adding random ties and not using data on the nodes' activities, the RT methods produce results of comparable quality. However, their costs in terms of conserved information is much higher (as we use the list of weights of nonbackbone links in the surrogate construction method).

VII. SUMMARY AND DISCUSSION

In this paper, we have considered how to bridge the gap between the backbone of a network and its actual use, particularly in data-driven numerical simulations of dynamic processes, i.e., how to turn network backbone extraction into the production of surrogate network data. Several backbone extraction procedures have indeed been put forward in the literature to extract a network's most important ties, which are supposed to summarize the most important information in the network. The issue of whether this summary suffices for actual data-driven uses has not been explored.

Here, we have tackled this issue for several types of backbones of a temporal network, by proposing systematic ways to construct surrogate data from the backbone information. We have then used these surrogate data in numerical simulations of epidemic processes and investigated how well the outcomes of simulations and the measure of epidemic risk match the simulations on the original data.

We have considered a wide variety of procedures, with different amounts and types of information on the original data kept in the summary of the original data formed by the backbone and completed by additional statistical information. The threshold backbone arbitrarily selects the links with the largest weights, while the significant tie filters are more principled and retain ties that cannot be explained by a null model. In all cases, the data summaries need to be informed by the number of temporal edges in the original data set. Still, the amount of other additional data they contain can vary significantly. In particular, these summaries might include information on the network structure and retain, or not, the values of the node activities computed on the whole data set. If these values are unknown, we have shown that it is possible to recompute approximate values by applying the ST filter null model to the backbone itself. Alternatively, it is possible to add ties at random between nodes to reach the original number of ties contained in the data, at the cost of also keeping the list of link weights of the nonbackbone links. The same procedure can be used to build surrogate data from the threshold backbone.

Most procedures yield surrogate data that allow us to obtain a reasonable approximation of the original outcome when used to simulate epidemic spreading processes. The quality of the approximation, however, depends on the surrogate's method. In particular, the information on the data's group structure turns out to play an important role, in line with other results showing its importance in diffusion processes [7,26]. Using realistic activity timelines of temporal edges also yields better results [7]. Once group information and realistic timelines are included in the construction of surrogate data, all methods give good results. The largest discrepancies between the original and surrogate data outcomes are obtained at large spreading and recovery parameters. This is not surprising as these parameters correspond to fast processes. In this case, the outcome can depend on the data's details [27] and temporal structures at short timescales that are not present in the surrogate data. For instance, in school data, temporal edges between classes occur in a synchronized way during the breaks [28], creating activity patterns that would need to be put by hand in the surrogate data and thus be contained in some way in the data summary.

Our results give hints on how to summarize complex data sets best so that they remain actionable. Moreover, as the construction of surrogate data is a stochastic process, each of the procedures discussed here yields an ensemble of surrogate data with similar statistical properties. This highlights an interesting potential application of our results. Indeed, collecting data sets is an expensive task, and several data properties depend on context, making modeling of realistic temporal networks a problematic task. Simultaneously, the availability of data with real properties is crucial to inform data-driven models of diffusion processes such as epidemics of infectious diseases. Moreover, collected data typically have a limited duration, and merely repeating the data might create undesired biases [22]. The various procedures we have described here make it possible to create synthetic surrogate data with properties very similar to empirical data without modeling assumptions. By tuning the backbone size, and hence the amount of surrogate data needed to be added to it, one can, moreover, tune the similarity between the original data and the surrogate replicas.

Our work has some limitations that also indicate the way for future work. First, we have limited our study to data describing contact between individuals. However, these data cover a broad range of contexts, have widely different temporal properties [21], and are particularly relevant for simulations of epidemic spread. Nevertheless, it would be interesting to more systematically study the dependence of our results on temporal structures. Second, we have considered only a limited number of backbone and surrogate construction methods. We sought to keep the methods parsimonious, so one could consider other backbone extraction methods, taking, e.g., temporal variations of the activities into account [15]. Finally, networks could support other types of processes, such as synchronization or complex contagion, which might also involve higher-order structures going beyond ties [29,30]. Correctly reproducing the outcome of such processes from a network summary might require the development of backbones of significant structures and corresponding new surrogate data construction methods.

ACKNOWLEDGMENTS

A.B. was supported by the ANR project DATAREDUX (ANR-19-CE46-0008) and JSPS KAKENHI Grant No. JP 20H04288. P.H. was supported by JSPS KAKENHI Grant No. JP 18H01655.

APPENDIX: DATA AND METHODS

1. Data

We use state-of-the-art publicly available data sets describing contacts between individuals in different settings, with high spatial and temporal resolution. All data were collected by the SocioPatterns collaboration, using an infrastructure based on wearable sensors that exchange radio packets, detecting close proximity ($\leqslant 1.5$ m) of individuals wearing the devices [23], with temporal resolution of 20 s. The data can be downloaded from the SocioPatterns website [17].

(i) The InVS15 data set contains the temporal network of contacts between individuals recorded in office buildings in France in 2015. The population is divided into 12 departments of varying sizes, but individuals of different departments can mix during the day with no time constraints [21].

(ii) The LyonSchool data set contains the contact events between 242 individuals (232 children and 10 teachers) in a primary school in Lyon, France, during 2 days in October 2009 [20]. The children are divided into ten classes of similar sizes (two classes per grade) and follow strict schedules, with mixing between classes limited to the breaks [20].

(iii) The Thiers13 data set gives the interactions between 327 students of nine classes of similar sizes within a high school in Marseille, during 5 days in December 2013 [19].

(iv) The SFHH conference data set describes the face-to-face interactions of 405 participants to the 2009 SFHH conference in Nice, France (4–5 June 2009) [21,22]. No metadata on the participants were collected and the resulting contact network does not show any group structure [22].

2. Significant ties backbones

For completeness, we recall here the procedure to extract the significant ties at a given significance level α from a temporal network [14].

We first define a temporal fitness model in which each node i has an activity level a_i, and the probability u that nodes i and j interact during any given time interval is given by the product of their activity levels, $u(a_i, a_j) = a_i a_j$.

Given a data set of N nodes and temporal length T, we estimate the node activity levels $\boldsymbol{a} \equiv (a_1^*, \dots, a_N^*)$ within the temporal fitness model from the N maximum likelihood equations

$$\sum_{j:j\neq i} \frac{m_{ij}^{\mathrm{o}} - T a_i^* a_j^*}{1 - a_i^* a_j^*} = 0, \ \forall\, i = 1, \dots, N, \tag{A1}$$

that can be solved by standard numerical algorithms. We then compute for each pair of nodes i and j the probability distribution of their total number of interactions m_{ij} in the null model, which is simply given by the binomial distribution

$$g(m_{ij}|a_i^*, a_j^*) = \binom{T}{m_{ij}} u(a_i^*, a_j^*)^{m_{ij}} [1 - u(a_i^*, a_j^*)]^{T - m_{ij}}. \tag{A2}$$

Let m_{ij}^c denote the cth percentile ($0 \leqslant c \leqslant 100$) of $g(m_{ij}|a_i^*, a_j^*)$. If the actual empirical number of interactions m_{ij}^{o} between i and j is larger than m_{ij}^c, it means that this empirical number cannot be explained by the null model at significance level $\alpha \equiv 1 - c/100$: in other words, i and j are connected by a significant tie at significant level α.

For a given value of α, the set of significant ties and the corresponding temporal edges form the ST backbone of the network. As α decreases, the number of significant ties obviously decreases, and one can tune α in order to obtain a backbone formed by a given fraction f of ties. Note that, as the significant ties tend to have a large number of interactions, the relative sizes of backbones in terms of number of temporal edges are higher than in terms of number of ties (see Table II).

When the nodes are divided into groups, we, moreover, consider a modified null model in which the probability of interaction at each time between i and j is

$$u_p(a_i, a_j) \equiv a_i a_j [\delta_{g_i, g_j} + p(1 - \delta_{g_i, g_j})], \tag{A3}$$

where g_i indicates the group of node i and δ is the Kronecker symbol.

For a given data set, we can write the maximum likelihood equations (MLEs) to estimate the node activity levels $\boldsymbol{a} \equiv (a_1^*, \dots, a_N^*)$ and the parameter p^*, similarly to the procedure of Ref. [14]: Given the null model, the number of times temporal edges are formed between nodes i and j over T time intervals is a random variable m_{ij} that follows a binomial distribution with parameters T and $u_p(a_i, a_j)$. Therefore, the joint probability function leads to

$$p(\{m_{ij}\}|\boldsymbol{a}, p) = \prod_{i,j:i\neq j} \binom{T}{m_{ij}} u_p(a_i, a_j)^{m_{ij}} [1 - u_p(a_i, a_j)]^{T - m_{ij}}, \tag{A4}$$

so the $N + 1$ MLEs are

$$\sum_{j:j\neq i} \frac{m_{ij}^{\mathrm{o}} - T a_i^* a_j^* [\delta_{g_i, g_j} + p^*(1 - \delta_{g_i, g_j})]}{1 - [\delta_{g_i, g_j} + p^*(1 - \delta_{g_i, g_j})] a_i^* a_j^*} = 0, \forall\, i = 1, \dots, N, \tag{A5a}$$

and

$$\sum_{i,j:g_i\neq g_j} \frac{m_{ij}^{\mathrm{o}} - T p^* a_i^* a_j^*}{1 - p^* a_i^* a_j^*} = 0. \tag{A5b}$$

The (G)ST filter can be applied to the original data set but also to the backbone itself. In Table III we give the correlation coefficients between the activities obtained by solving the MLE equations for a data set and for its extracted backbone representing a fraction f of ties.

3. Surrogate data

As described in the main text, we have put forward several methods to build surrogate data starting from a backbone. These methods consist of two steps, first choosing the surrogate ties and then creating timelines of temporal edges on each tie.

In Table IV, we summarize each method's main points for each type of backbone, the data needed in addition to the backbone information, and the size of these additional data. Note that the random links methods need several inputs of the order of the number of ties in the original data, while the methods based on the null model instead use an input scaling with the number of nodes. The method needing the fewest extra data is the one recomputing the activities applying the ST filter methodology on the backbone data itself.

In the methods based on the (G)ST null models, we need to calibrate the parameter α (or the two parameters α_{intra} and α_{inter}). To this aim, we first try at each time stamp to add a temporal edge with probability $a_i a_j$ for each (i, j) not in the backbone. This creates a total number of temporal edges E_T'. The actual number of surrogate temporal edges needed is actually $E_T - E_{bT}$, i.e., the difference between the number

TABLE IV. Summary of the various methods to choose the surrogate ties.

Backbone type	Surrogate type	Method summary	Extra data needed	Size of extra data needed
ST	ST-OA	For each (i, j) not in backbone, at each time stamp add a temporal edge with probability $\alpha a_i a_j$, with α scaled to adjust the total number of temporal edges	List of original node activities; Number of temporal edges	$N + 1$
	ST-RA	Compute the activity $\tilde{a}_i$ of each node with the ST backbone method applied on the backbone itself; add surrogate ties as for the ST-OA method, using the recomputed activities	Number of temporal edges	1
	ST-RT	Add ties at random in order to reach the number of ties of the original data, with weights extracted at random from the list of weights of the nonbackbone ties	Number of ties; list of weights of ties not in backbone	$E - E_b + 1$
TB	TB-RT	Same as for ST-RL	Number of ties; list of weights of ties not in backbone	$E - E_b + 1$
GST	GST-OA	For each (i, j) not in backbone, at each time stamp add a temporal edge with probability $\alpha_{\text{intra}} a_i a_j$ if i and j are in the same group, $\alpha_{\text{inter}} p a_i a_j$ else, with $\alpha_{\text{intra/inter}}$ scaled to adjust the total number of temporal edges within and between groups	List of original node activities and parameter p; group membership; number of temporal edges within groups and between groups	$2N + 3$
	GST-RA	Compute the activity $\tilde{a}_i$ of each node and the parameter $\tilde{p}$ with the GST backbone method applied on the backbone itself; add surrogate ties as for the GST-OA method, using the recomputed activities	Group membership; number of temporal edges within groups and between groups	$N + 2$
	GST-RT	Add ties at random in order to reach the same number of ties within and between groups as in the original data, with weights extracted at random from the list of weights of the nonbackbone ties	Group membership; number of ties between and within groups; list of weights of nonbackbone ties, between and within groups	$N + E - E_b + 2$
TB	GTB-RT	Same as for GST-RT	Group membership;number of ties between and within groups; list of weights of nonbackbone ties, between and within groups	$N + E - E_b + 2$

of temporal edges in the data and in the backbone. Therefore, we set $\alpha = (E_T - E_{bT})/E'_T$ and we use as probabilities of creation of temporal edges $\alpha a_i a_j$. When the data group structure is taken into account, the procedure is performed separately for intra- and intergroup ties. We note that the final number of temporal edges in the surrogate data is not strictly fixed by this procedure but remains a stochastic outcome. The number of ties is not fixed either but is also an outcome of the procedure, contrary to the procedures based on adding random links.

Finally, to construct surrogate timelines respecting the data statistics of event and interevent durations, we proceed as follows, for each tie (i, j) with number of temporal edges n_{ij}:

(1) We extract a random initial time t_0 in $[0, T]$; all the times are then considered modulo T; we set $n = n_{ij}$.

(2) We iterate the following procedure until $n = 0$, i.e., until n_{ij} temporal edges have been created:

(a) Extract a random duration τ from the fitted distribution of event durations.

(b) Check that $\tau \leqslant n$, else replace τ by n.

(c) Add τ temporal edges between i and j, namely, on the interval $[t_0, t_0 + \tau - 1]$.

(d) Extract a random interevent time Δt from the fitted distribution of interevent durations.

(e) Replace t_0 by $t_0 + \tau + \Delta t$ and n by $n - \tau$.

4. Simulations of the epidemic spread

For the simulation of the SIR model on temporal networks we use the approach and code presented in Ref. [31]. We start the simulation with all nodes susceptible and introduce the disease at a random node at a random time (uniformly chosen between the beginning and end of the temporal network). Then if there is an event between a susceptible and an infectious, a contagion occurs with a probability β. The infected person recovers with a rate ν; i.e., the time to recovery is a random variable δ extracted from the distribution $\nu \exp(-\nu\delta)$ [31]. Finally, we assume that an individual that gets infected at time t' cannot infect anyone else until $t > t'$. For every pair of parameter values (β, ν), we run this algorithm 10^7 times for averages.

We calculate the basic reproductive number R_0 directly from the simulations as the average numbers of individuals infected directly by the source. Calculating the average outbreak size Ω is a similarly straightforward average over the number of nodes in the recovered state when the outbreak is extinct. If the outbreak is not extinct when the simulation reaches the end of the data set, the outbreak size is the number of nodes in either the infectious or the recovered state at the last time stamp of the data.

[1] R. Albert and A.-L. Barabási, Rev. Mod. Phys. **74**, 47 (2002).

[2] A. Barrat, M. Barthélemy, and A. Vespignani, *Dynamical Processes on Complex Networks* (Cambridge University Press, Cambridge, UK, 2008).

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

[4] P. Holme, Phys. Rev. E **94**, 022305 (2016).

[5] N. Masuda and R. Lambiotte, *A Guide to Temporal Networks* (World Scientific, Singapore, 2016).

[6] Y. Liu, T. Safavi, A. Dighe, and D. Koutra, ACM Comput. Surv. (CSUR) **51**, 62 (2018).

[7] M. Génois, C. L. Vestergaard, C. Cattuto, and A. Barrat, Nat. Commun. **6**, 8860 (2015).

[8] J. Fournet and A. Barrat, Sci. Rep. **7**, 9975 (2017).

[9] M. Á. Serrano, M. Boguná, and A. Vespignani, Proc. Natl. Acad. Sci. USA **106**, 6483 (2009).

[10] G. Casiraghi, V. Nanumyan, I. Scholtes, and F. Schweitzer, in *International Conference on Social Informatics* (Springer, Berlin, 2017), pp. 111–120.

[11] M. Tumminello, S. Micciché, F. Lillo, J. Piilo, and R. N. Mantegna, PLoS One **6**, e17994 (2011).

[12] V. Hatzopoulos, G. Iori, R. N. Mantegna, S. Miccichè, and M. Tumminello, Quant. Finance **15**, 693 (2015).

[13] L. M. Shekhtman, J. P. Bagrow, and D. Brockmann, J. Complex Networks **2**, 110 (2014).

[14] T. Kobayashi, T. Takaguchi, and A. Barrat, Nat. Commun. **10**, 220 (2019).

[15] M. Nadini, C. Bongiorno, A. Rizzo, and M. Porfiri, Nonlinear Dyn. **99**, 855 (2020).

[16] M. Nadini, A. Rizzo, and M. Porfiri, J. Phys.: Complexity **1**, 015001 (2020).

[17] SocioPatterns, http://www.sociopatterns.org.

[18] M. Génois, C. L. Vestergaard, J. Fournet, A. Panisson, I. Bonmarin, and A. Barrat, Network Sci. **3**, 326 (2015).

[19] R. Mastrandrea, J. Fournet, and A. Barrat, PLoS One **10**, e0136497 (2015).

[20] J. Stehlé, N. Voirin, A. Barrat, C. Cattuto, L. Isella, J. Pinton, M. Quaggiotto, W. Van den Broeck, C. Régis, B. Lina *et al.*, PLoS One **6**, e23176 (2011).

[21] M. Génois and A. Barrat, EPJ Data Sci. **7**, 11 (2018).

[22] J. Stehlé, N. Voirin, A. Barrat, C. Cattuto, V. Colizza, L. Isella, C. Régis, J.-F. Pinton, N. Khanafer, W. Van den Broeck *et al.*, BMC Med. **9**, 87 (2011).

[23] C. Cattuto, W. Van den Broeck, A. Barrat, V. Colizza, J.-F. Pinton, and A. Vespignani, PLoS One **5**, e11596 (2010).

[24] A. Machens, F. Gesualdo, C. Rizzo, A. E. Tozzi, A. Barrat, and C. Cattuto, BMC Infect. Dis. **13**, 185 (2013).

[25] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevE.103.052304 for supplementary results.

[26] T. Smieszek and M. Salathé, BMC Med. **11**, 35 (2013).

[27] L. Isella, J. Stehlé, A. Barrat, C. Cattuto, J.-F. Pinton, and W. Van den Broeck, J. Theor. Biol. **271**, 166 (2011).

[28] L. Gauvin, A. Panisson, and C. Cattuto, PLoS One **9**, e86028 (2014).

[29] I. Iacopini, G. Petri, A. Barrat, and V. Latora, Nat. Commun. **10**, 2485 (2019).

[30] F. Battiston, G. Cencetti, I. Iacopini, V. Latora, M. Lucas, A. Patania, J.-G. Young, and G. Petri, Phys. Rep. **874**, 1 (2020).

[31] P. Holme, PLoS One **16**, e0246961 (2021).