Estimating the outcome of spreading processes on networks with incomplete information: A dimensionality reduction approach

Anna Sapienza
Information Sciences Institute, Viterbi School of Engineering, University of Southern California, Marina del Rey, California 90292, USA and Data Science Laboratory, ISI Foundation, 10126 Turin, Italy

Alain Barrat
Aix Marseille Univ, Université de Toulon, CNRS, CPT, Marseille, France and Data Science Laboratory, ISI Foundation, Turin, Italy

Ciro Cattuto and Laetitia Gauvin
Data Science Laboratory, ISI Foundation, 10126 Turin, Italy

(Received 4 October 2017; revised manuscript received 27 March 2018; published 30 July 2018)

Recent advances in data collection have facilitated the access to time-resolved human proximity data that can conveniently be represented as temporal networks of contacts between individuals. While the structural and dynamical information revealed by this type of data is fundamental to investigate how information or diseases propagate in a population, data often suffer from incompleteness, which possibly leads to biased estimations in data-driven models. A major challenge is thus to estimate the outcome of spreading processes occurring on temporal networks built from partial information. To cope with this problem, we devise an approach based on non-negative tensor factorization, a dimensionality reduction technique from multilinear algebra. The key idea is to learn a low-dimensional representation of the temporal network built from partial information and to use it to construct a surrogate network similar to the complete original network. To test our method, we consider several human-proximity networks, on which we perform resampling experiments to simulate a loss of data. Using our approach on the resulting partial networks, we build a surrogate version of the complete network for each. We then compare the outcome of a spreading process on the complete networks (nonaltered by a loss of data) and on the surrogate networks. We observe that the epidemic sizes obtained using the surrogate networks are in good agreement with those measured on the complete networks. Finally, we propose an extension of our framework that can leverage additional data, when available, to improve the surrogate network when the data loss is particularly large.

DOI: 10.1103/PhysRevE.98.012317

I. INTRODUCTION

The advances made in data collection technologies have led to a wealth of high-resolution time-resolved data. Mobile sensing devices, social networking applications, and wearable sensors have indeed significantly contributed to monitor social interactions and physical proximity of individuals in time [1–6]. Such fine-grained data monitoring is crucial for a deeper study of human proximity dynamics, described by complex temporal networks, in which links are drawn between nodes representing individuals when they are in close range [7], and their interplay with contagion processes. Physical proximity interactions indeed play a fundamental role in conveying information or in the spread of diseases [8–11]. They can thus inform our understanding of how messages or infectious diseases such as flulike illnesses propagate among individuals.

However, despite the efforts made to increase the accuracy in the data collection, relational data often suffer from incompleteness, resulting in missing links in empirical networks [12]. This lack of information can arise for several reasons: limited participation during surveys, incomplete records (diary based or device based) [13–16], and technical issues occurring during the data collection process. In the case of smartphone sensing such as in [3,5], proximity might be undetected during some time windows. For instance, people might turn off their Bluetooth or call detail records might not provide access to the vicinity of individuals to a cell tower at each time, possibly leading to undetected co-presence events. In the case of self-reporting of sexual relationships, individuals might choose not to disclose all of their partners, which leads to biases when focusing on the spread of sexually transmitted disease [17]. Data incompleteness not only can affect the measured properties and structure of temporal networks [18,19] but can also reflect on the simulated evolutions of contagion processes, leading to inaccurate conclusions [17,20,21]. The investigation of information or disease propagation processes in data-driven models built using such data must thus be undertaken carefully.

Several approaches have been put forward to recover missing links in networks [22]. These methods include distributional models, which estimate the likelihood of the presence of a link on the basis of the observed links and nodes attributes [23], hierarchical structure methods [24], stochastic block models [25,26], and expectation maximization methods [27],

which try to extract the connectivity patterns in the available part of the network to infer and complete the unknown part. The goal of these methods is the exact recovery of the missing links, a complicated task that becomes nearly impossible to achieve when a large amount of data is missing. Notably, this might actually not be necessary if the goal is to estimate the global outcome of a contagion process at the population level and not the risk concerning a specific individual. Starting from this point of view, several approaches have thus been developed to specifically estimate important properties of epidemic spread and information cascade without trying to recover the original network [15,20,28–30]. These methods, however, either are process specific or rely in a fundamental way on the existence of metadata (allowing one to define groups in the population) in the network, together with the knowledge of the structure to which each population member belongs.

Here we propose a self-contained approach that does not rely on metadata and addresses the issue of missing data directly for temporal networks, while most existing methods are based on static network representations. Moreover, our approach is not based on a microscopic scale, i.e., on statistical distributions of local quantities, or on macroscopic properties of the network, but on an intermediate scale (which we refer to as the mesoscale level), i.e., modular structures typically observed in networks that are relevant for spreading processes [31]. As in [15,20], the aim of this work is not to recover the exact links which are missing in the data but instead to build a surrogate version of the network of interactions. The main difference with these previous works, which is a crucial building block of our approach, is that we uncover *in the incomplete data* latent structures that involve nodes affected by missing activity.

To study the network at this mesoscale level, we take advantage of tensor decomposition techniques [32] to extract both the topological and temporal properties of the network [33]. Without using metadata or external information about the nodes, our method allows us to recover fundamental properties of the studied temporal network, such as the temporal activity of nodes with partial information (i.e., the temporal evolution of their number of contacts). Differently from other methods that apply tensor factorization techniques for tensor completion and prediction [34–36], we use this approach to build a surrogate version of the network that yields a correct estimate of the outcome of a simulated spreading process occurring on the network. Furthermore, we show that the framework presented can naturally be extended to take advantage of other sources of information, correlated with the original network, that might be available, such as information deriving from subsidiary data sources. In particular, we use approximated location data as a subsidiary source for contact data, as two individuals must be in the same location to be in contact, thus making the location data correlated with the contacts.

The paper is organized as follows. In Sec. II we present the notation used in the paper. In Sec. III we describe the problem statement. In Sec. IV we explain the method that we developed to carry out the study. In Sec. V we report the results achieved by our approach in the study of several temporal human proximity networks. In Sec. VI we discuss the performance and limitations of the method and future research directions.

II. NOTATION

The following notation is used throughout the present paper. Lowercase letters denote scalar variables, e.g., t, capital letters denote defined constants, e.g., T, and boldface lowercase letters denote vectors, e.g., $\mathbf{t}$. Matrices are denoted by boldface capital letters, e.g., $\mathbf{T}$, where the ith column of a matrix $\mathbf{T}$ is $\mathbf{t}_i$ and the (i,j) entry is t_{ij}. Third-order tensors are denoted by bold calligraphic letters, e.g., $\mathcal{T}$, whose (i,j,k) entry is t_{ijk}. The tensor product is denoted by $\circ$, the Hadamard (elementwise) product by $*$, the Kronecker product by $\otimes$, and the outer product by $\cdot$. The Frobenius norm is denoted by $\|\cdot\|_F$.

III. PROBLEM STATEMENT

In this work, we aim at reproducing the outcome of contagion processes on temporal networks of human interactions, by starting from incomplete information on these networks. To this aim, we consider a scenario in which part of the activity of a fraction of the network nodes (i.e., part of their interactions over time) is missing in the data and we only assume to know which nodes might be affected by missing data.

To provide an estimate of the outcome of spreading processes on the network, we do not try to recover the exact missing links and interaction events of these nodes. Our method aims instead at building a surrogate version of the complete network by taking advantage of the only information that is available. As we will describe in detail in Secs. IV A and IV B, this information can be related either only to the partial network of human contacts or to richer data that can be used as a proxy for human proximity. For instance, we could have access both to the partial temporal contact network and to the approximated location of individuals provided by smartphones, through GPS, Bluetooth, or WiFi signals: Such additional information can conveniently be represented, in the same spirit as in [37], as a temporal bipartite network between individuals and locations, in which a link is drawn between an individual and a location when the individual is detected in that location. We will show how such information can be integrated in our framework.

The underlying assumption of our method is that we can leverage the mesoscale properties of the partial network, such as the presence of correlations in the node activities, to build a surrogate version of the complete network. To extract these topological and temporal properties from the incomplete data we rely on the non-negative tensor factorization (NTF) technique [32]. In particular, our method is based on a NTF framework handling missing values [38,39]. By applying the NTF on a tensor representing a temporal network, we can indeed identify groups of nodes having similar connectivity patterns and whose links have similar activation times. Each of these groups can be seen as a subnetwork. Note that a latent structure is defined by nodes sharing some links and having correlated activity. By construction, such a structure has to be composed by at least three nodes to form at least two links which co-occur at least once. By studying the structure and the temporal activity of each subnetwork, we can infer the properties of the nodes whose activity is partially missing. Our method is thus divided into two main steps that we detail in the next section. The first step consists in extracting latent structures from the partial temporal network through NTF

adapted to handle incomplete information. The second step is the construction of a surrogate network.

IV. METHOD

A. Extracting latent structures from a partial temporal network

Three-way tensors (i.e., three-dimensional arrays) are natural representations of temporal networks: Given an undirected temporal network, composed of N nodes and $k = 1, \ldots, K$ time intervals, we can represent its snapshots $G_k = (V, E_k)$, which have $|V| = N$ nodes and a set of links E_k, by K adjacency matrices $\mathbf{M}_k \in \mathbb{R}^{N\times N}$ of the form

$$\mathbf{M}_k = \begin{cases} m_{ij} = 1 & \text{if } (i,j) \in E_k \\ m_{ij} = 0 & \text{otherwise.} \end{cases}$$

These adjacency matrices form the slices of a tensor $\mathcal{T} \in \mathbb{R}^{I\times J\times K}$, where $I = J = N$. In the case of missing data, zero values in the tensor can correspond either to no activity or to undetected activity. We need to factorize however only the part of the tensor that corresponds either to measured activity or to actual inactivity (absence of contact). To this aim, the zero entries that correspond to possibly undetected activities have to be masked in the tensor. If a node with partial information has no activity at all measured during a given snapshot (in a given slice of the tensor), we consider the related zero entries in the tensor $\mathcal{T}$ as possible undetected activities, i.e., possible missing contacts. On the other hand, if a node, for which we know that some data are missing, has at least one contact measured in a slice, we assume that no information about that node was lost at all in that tensor slice. In other terms, we assume for each node and each time slice that either all or none of the activity recorded by that node in that time slice is present in the data. For instance, this would reflect the case in which a sensor measuring proximity is turned off or the GPS coordinates of a mobile user are not collected during a time window: All the proximity information concerning this sensor in this time window is lost. However, if the measuring device is not turned off, all the proximity relationships with other devices that are also on during that time window are present in the incomplete data. We explain at the end of this section the possible impact of the choice of such a hypothesis.

We thus introduce a binary tensor $\mathcal{W}$, of the same size as $\mathcal{T}$, whose entries are defined as

$$w_{ijk} = \begin{cases} 0 & \text{if } i \text{ or } j \text{ has no activity in the time window } k \\ 1 & \text{otherwise.} \end{cases}$$

This tensor is used to mask the part of the tensor $\mathcal{T}$ that might be linked to possibly undetected activity. The approximation of the masked tensor $\mathcal{T} * \mathcal{W}$ consists in minimizing the cost function with non-negative constraints [39]

$$f_w(\boldsymbol{\lambda}, \mathbf{A}, \mathbf{B}, \mathbf{C}) = \left\| \mathcal{W} * \left(\mathcal{T} - \sum_{r=1}^{R} \lambda_r \mathbf{a}_r \cdot \mathbf{b}_r \cdot \mathbf{c}_r \right) \right\|_F^2,$$

where R is the rank of the approximation and is hereinafter called the number of components. The vectors $\mathbf{a}_r$, $\mathbf{b}_r$, and $\mathbf{c}_r$ with $r \in [1, R]$ form, respectively, the factor matrices $\mathbf{A} \in \mathbb{R}^{I\times R}$, $\mathbf{B} \in \mathbb{R}^{J\times R}$, and $\mathbf{C} \in \mathbb{R}^{K\times R}$.

Each component, i.e., each tuple of vectors $(\mathbf{a}_r, \mathbf{b}_r, \mathbf{c}_r)$, is interpreted as a latent structure in the data set: The sum $\mathcal{T}_{\text{app}} = \sum_{r=1}^{R} \lambda_r \mathbf{a}_r \cdot \mathbf{b}_r \cdot \mathbf{c}_r$ yields thus an approximation of $\mathcal{T}$ that can be interpreted as a superposition of R latent structures. The vectors $\mathbf{a}_r$ and $\mathbf{b}_r$ indicate which nodes and links participate in the latent structure r, while $\mathbf{c}_r$ describes the temporal activity of the structure r. Let us note that here we impose that the vectors $\mathbf{a}_r$, $\mathbf{b}_r$, and $\mathbf{c}_r$ are normalized, which is made possible by the use of the scalar parameters λ_r. This is an important point for building the surrogate network as described below. For the sake of readability, the cost function is rewritten in the following form:

$$f_w(\boldsymbol{\lambda}, \mathbf{A}, \mathbf{B}, \mathbf{C}) = \|\mathcal{W} * (\mathcal{T} - [[\boldsymbol{\lambda}; \mathbf{A}, \mathbf{B}, \mathbf{C}]])\|_F^2. \quad (1)$$

Using the mask $\mathcal{W}$ amounts to approximating the tensor $\mathcal{T}$ based only on information of which we are certain: We are approximating only the part of the tensor composed of elements that are either 1, corresponding to measured events, or 0, corresponding to the real absence of contact, without taking into account all the 0 values that might correspond to undetected activity.

In other terms, minimizing the cost function f_w corresponds to finding the best approximation of the nonmasked part of the tensor by a sum of components, each corresponding to a latent structure. The analysis of $\mathbf{A}$, $\mathbf{B}$, and $\mathbf{C}$ yields then information on which link is involved in which mesoscale structure, including in particular the links involving nodes with incomplete information. Several methods are available to estimate the factor matrices $\mathbf{A}$, $\mathbf{B}$, and $\mathbf{C}$ [40,41] and details are given in Sec. VII.

B. Extracting latent structures from coupled temporal networks

Here we propose an extension of the latent structure detection method to the case in which we have access to richer information, not expressible in one single temporal network. The integration of such information might help to better recover the missing entries in the tensor describing the temporal contact network: For instance, if we have access, in addition to the partial contact network, to the location of individuals involved in the contacts, the latter can contribute to recover the missing information on contacts. To integrate such additional data, we propose to use the so-called joint non-negative tensor factorization (JNTF) [42] that makes it possible to decompose multiple temporal networks at once in a coupled manner (in practice, we will consider the partial temporal contact network and a position network evolving with time).

Let us consider the general case of S different temporal networks, each represented by a tensor $\mathcal{T}_s$, with possibly different dimensions as they might represent different types of information. We can approximate them in a coupled way by computing the following minimization problem with non-negative constraints:

$$\min \sum_{s=1}^{S} \|\mathcal{T}_s - [[\boldsymbol{\lambda}_s; \mathbf{A}_s, \mathbf{B}_s, \mathbf{C}_s]]\|_F^2 \text{such that } \boldsymbol{\lambda}_s, \mathbf{A}_s, \mathbf{B}_s, \mathbf{C}_s \geqslant 0. \quad (2)$$

Different couplings can be considered by imposing that some of the factor matrices $\mathbf{A}_s$, $\mathbf{B}_s$, and $\mathbf{C}_s$ in the equation are equal for different values of s. The idea behind the introduction of

this coupling is that different networks can provide partially redundant information and that this redundancy is relevant for recovering missing entries. For instance, if we have access to the locations of nodes but only to partial information concerning their contacts, we can couple the decomposition of the tensor $\mathcal{T}_1$ representing the partial contact network to the decomposition of the tensor $\mathcal{T}_2$ representing the time evolution of the location of nodes and gain in this way information on the possible contacts over time. In practice, we would impose $\mathbf{A}_1 = \mathbf{A}_2$ and $\mathbf{C}_1 = \mathbf{C}_2$, which correspond, respectively, to imposing the same nodes' memberships and the same activity timeline for each latent structure in the resulting approximations of the contact and location tensors.[1]

The joint factorization of tensors, including one with missing information $\mathcal{T}_{s'}$, can be adapted to handle missing values in the same way as the NTF:

$$\min \left(\alpha_{s'} \|\mathcal{W} * (\mathcal{T}_{s'} - [[\boldsymbol{\lambda}_{s'}; \mathbf{A}_{s'}, \mathbf{B}_{s'}, \mathbf{C}_{s'}]])\|_F^2 + \sum_{\substack{s=1 \\ s \neq s'}}^{S} \alpha_s \|\mathcal{T}_s - [[\boldsymbol{\lambda}_s; \mathbf{A}_s, \mathbf{B}_s, \mathbf{C}_s]]\|_F^2 \right)$$
$$\text{such that } \boldsymbol{\lambda}_s, \mathbf{A}_s, \mathbf{B}_s, \mathbf{C}_s \geqslant 0 \, \forall s.$$

To solve this minimization problem we adapted the active-set-like method with Karush-Kuhn-Tucker optimality conditions [40,43] (see Sec. VII for details). In the following, we set all the α_s parameters equal to 1. We note, however, that they could be used to tune the relevance of the information provided by each tensor. As an example, by setting $\alpha_{s'} > \alpha_s \, \forall s \neq s'$, we would give more importance to the information provided by the network with partial information than to the one given by the subsidiary data.

C. Surrogate network

As introduced in the previous sections, both types of factorizations, either with or without the use of auxiliary data, approximate the partial tensor (network) as $\mathcal{T}_{\text{app}}$, which is fully defined by the three factor matrices $\mathbf{A}$, $\mathbf{B}$, and $\mathbf{C}$ and the vector λ. The columns of the factor matrices correspond to the latent structures, which can be interpreted as subnetworks whose links have similar temporal properties, as we now describe.

Each tuple $(\mathbf{a}_r, \mathbf{b}_r, \mathbf{c}_r)$ obtained by the factorization step gives information on the level of membership of each link for the corresponding structure and on the times in which these links are active in the network (i.e., on the level of activation of the links). First, we obtain the membership of each link (i, j) to each structure r by using $\mathbf{A}$ and $\mathbf{B}$; this membership is indeed given by the (i, j)th element of the Kronecker product

$$\mathbf{a}_r \otimes \mathbf{b}_r = \mathbf{a}_r \cdot \mathbf{b}_r^T.$$

As we consider undirected networks, the actual membership of each link is symmetrized in the following way:

$$\frac{\mathbf{a}_r \cdot \mathbf{b}_r^T + \mathbf{b}_r \cdot \mathbf{a}_r^T}{2}.$$

For each r, we rank these values in decreasing order and we consider that the links with the largest membership values, composing 95% of the total sum of the squared memberships, belong to the component r (note that, depending on the membership values, a link could belong to more than one component). Then, to detect the times in which each latent structure is active, we consider the matrix $\mathbf{C}$, which summarizes the temporal activity of each component; by using the Otsu method [44], a common way to perform binary thresholding, we transform the temporal activity of each component in its binary version (1 if it is active, 0 otherwise). Note that this step is possible because we work with normalized vectors $\mathbf{c}_r$.

For each structure, we then select the links involving at least one node with partial information and the times in which its activity was potentially lost (in the sense that no activity of that node was recorded at all for this particular time) and we add the corresponding elements of the structure to the partial tensor $\mathcal{T}$ to create the surrogate tensor $\mathcal{T}_{\text{surrogate}}$. The rationale is that the factorization has allowed us to determine, for the links for which only partial information is available, to which latent structures they belong. Using the activity timelines of the latent structures, we thus reconstruct the missing parts of the activity timelines of these links. The procedure of binarization of the approximated tensor described above is summarized in Fig. 1.

To find the surrogate tensor by using the JNTF, an addition is necessary. The JNTF indeed yields components influenced by the activity in both the partial proximity network and the subsidiary temporal network (such as a network linking individuals and locations). Thus, the JNTF will approximate the network in a way based also on co-presence events of individuals. As co-presence is a necessary but not sufficient condition for two individuals to be in contact, the links added when creating $\mathcal{T}_{\text{app}}$ are less likely to correspond exactly to contacts that were missing than in the NTF case. This results in link weights larger than in the original data, with less bursty activity patterns (the weight of a link is given by the number of time slices in which it is active). As mentioned in Sec. IV B, the weight of the information provided by the second tensor can of course be tuned in the JNTF by adding coefficients so that the components extracted by the factorization are more representative of the network with partial information than of the other one. The investigation of this possibility, however, is outside the scope of this paper. The procedure of binarization of $\mathcal{T}_{\text{app}}$ described above needs thus to be completed to make the properties of the network more heterogeneous. In order to do this, we have to discard some of the times in which the links involving nodes with partial information are active in $\mathcal{T}_{\text{app}}$ to compensate for the overestimation and get closer to the empirical network. The key idea here is to zero out some elements in $\mathcal{T}_{\text{surrogate}}$ to recover a distribution of link weights comparable to the empirical one, measured on the links involving only nodes with no missing information. We pick from this distribution a number of weights equal to the number of links having partial information, for which we

[1] Note that because of the auxiliary information the factor matrices $\mathbf{A}_1 = \mathbf{A}_2$, $\mathbf{B}_1$, and $\mathbf{C}_1 = \mathbf{C}_2$ provided by the JNTF differ from those found when applying the NTF.

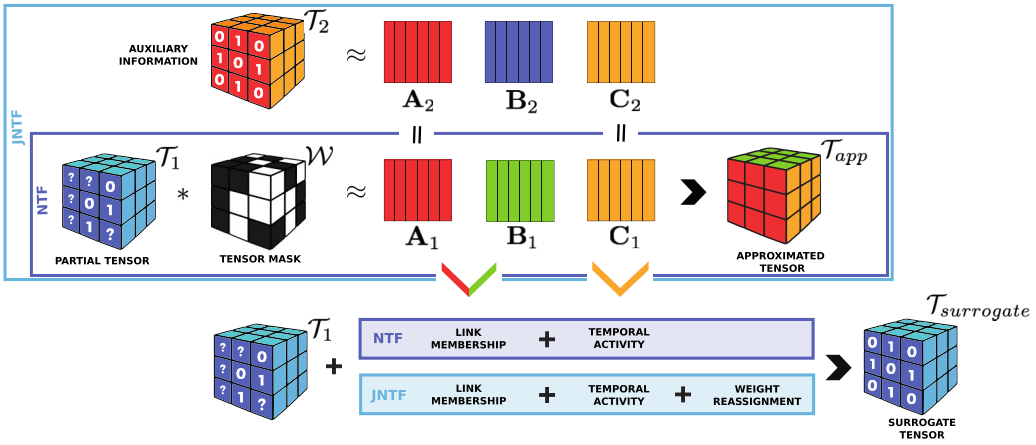

FIG. 1. Schematic representation of the approach for creating the surrogate network in the case in which we just have access to a partial network, such as a temporal contact network (NTF), and in the case in which we also have access to additional data, such as the approximate location of the nodes (JNTF). The main steps are (i) masking of the tensor in order to retain only the part of the data we are sure about (for each link, the times at which we know if it is active or not), (ii) factorization that approximates the tensor as a sum of components interpreted as latent structures, and (iii) extraction of link memberships and temporal activity for each latent structure. Finally, we use this information to complete the partial tensor and obtain the surrogate network. In order to exploit potential additional data, the factorization step can be done in a joint way (JNTF) where we constrain the factor matrices related to the common dimensions to be equal ($\mathbf{A}_1 = \mathbf{A}_2$ and $\mathbf{C}_1 = \mathbf{C}_2$); in this case, an additional step of weight reassignment is needed.

have to adjust the weights, and we assign them at random to these links. Finally, we compare for each link the new value to its weight in $\mathcal{T}_{\text{surrogate}}$: If the new weight is smaller than the old one, we erase at random parts of the link activity that are present in $\mathcal{T}_{\text{surrogate}}$, until we reach the new weight; if instead the new weight is larger than the old one, we do not act on that link's activity. The reason why we can rely on the weight distribution measured on the partial network is due to its robustness to sampling, as shown in [4,15,20,28] for various sampling procedures and in Appendix C for the sampling considered here. After the reassignment of weights, the resulting tensor $\mathcal{T}_{\text{surrogate}}$ is used to approximate the whole temporal network of contacts and perform simulations of spreading processes. The procedure in the case of additional data is also summarized in Fig. 1.

D. Summary outline of the approach

The steps of our method can be summarized as follows.

(i) Given a tensor with missing data, representing a temporal network (weighted or unweighted), the NTF approximates such a tensor through the sum of R components, interpreted as structures. Each component is fully determined by three normalized vectors $\mathbf{a_r}$, $\mathbf{b_r}$, and $\mathbf{c_r}$ and a scalar λ_r.

(ii) We use the vectors $\mathbf{a_r}$ and $\mathbf{b_r}$ to identify the links involved in each component r. Once the list of links is identified, we further select those involving at least a node with partial information.

(iii) We need now to assign to these links their temporal activation and reconstruct the surrogate tensor. To this aim, we binarize the vector $\mathbf{c_r}$ ($\mathbf{c_r}$ is normalized and its elements are bounded between 0 and 1) by using the Otsu method. The binarization provides the times at which the links should be present.

(iv) At this stage we have obtained the needed information about the entries t_{ijk} to insert back in the incomplete tensor: the identified links (i,j) at the selected times k. In the case study we just insert in the tensor a new value equal to 1 (as the original network is unweighted). Note that this could be extended to the case of a weighted network, as we could then insert back the approximated value of T_{app} at position (i,j,k). This value is given by the product of a_{ri}, b_{rj}, c_{rk}, and λ_r.

An additional step of weight reassignment is needed in the case of the joint non-negative factorization, as explained previously.

Let us note that we consider here a type of data loss that would correspond to having devices turned off at some moments: For each node and at each time slice, either all or none of the activity recorded by that node in that time slice is present in the data. This assumption comes into play in two ways. (i) In the construction of the mask $\mathcal{W}$, the assumption allows us to make the approximation by a sum of latent structures using the information on both activity times and inactivity times of the links. It would be possible to relax this assumption by using only the values of 1 in $\mathcal{T}$, i.e., the times at which links are known to be active. The tensor $\mathcal{W}$ would then be used to mask all the 0 entries t_{ijk} in the incomplete tensor, where either i or j is a node affected by data loss. (ii) When we are adding the recovered links (involving nodes with missing information) to the partial tensor, we indeed add back links active only at times in which the involved node with partial information had no interaction at all. To relax this assumption, we could add back links involving nodes with partial information at times chosen according to a criterion on the node activity, for instance, if the activity of the nodes with missing information at that times is below the average activity of nodes without missing information.

E. Baseline network

In order to have a reference to which to compare our approach, we additionally consider a baseline procedure that

assigns activities to nodes with partial information based simply on the average activity of the other nodes. In this procedure, we first compute the average total activity of the nodes for which no information is missing, a_{av}; the total activity of a node is simply given by its total number of contacts. We also compute the total activity for each node with potentially missing information; if it is lower than a_{av}, we add contacts at times chosen at random among the times in which at least one other node had some activity (in order to avoid adding contacts in periods with no activity at all, such as the nights in contact network data), until the activity of the node reaches a_{av}. We obtain in this way our baseline surrogate network.

V. RESULTS

We apply the method described in the previous sections to three temporal human proximity networks. The data were collected by the SocioPatterns collaboration [45] in two conferences in Italy (HT09) and France (SFHH) and in a primary school in France (LSCH). In each case, the proximity of individuals in the network was measured with wearable sensors able to capture face-to-face contacts occurring in a (1–2)-m range with a 20-s time resolution. As for the purpose of the present paper we do not need such a high resolution, we aggregate the data to a 15-min resolution. The HT09 data set was collected during the conference in [46]. The resulting temporal network has $N = 113$ nodes and $K = 237$ snapshots. The SFHH data set was collected during the conference of the Société Française d'Hygiène Hospitalière, yielding a temporal network with $N = 417$ nodes and $K = 129$ time snapshots [4]. Finally, the LSCH data set was collected in a primary school in Lyon [47]. The resulting network has $N = 241$ nodes and $K = 130$ snapshots. We consider all the snapshots as unweighted networks (meaning that each element of the tensor is either 0 or 1).

We simulate on each data set a loss of data determined by a fraction of nodes $p_{\text{nodes}} \in [0.1,0.2,0.4]$ for which we zero out the activity occurring during the first half of the total time span, i.e., the information concerning each of these nodes is lost over a fraction $p_{\text{times}} = 0.5$ of the temporal snapshots. We consider in scenarios where all the nodes are affected by the data loss at the same times and in particular in the first part of the time range. This choice corresponds indeed to a worst-case scenario maximizing the impact of the loss of information on the spreading process. To show the versatility of the method, we also consider in the Appendix D scenarios in which data losses occur at random times for each affected node.

In practice, for each data set we start from the complete tensor $\hat{\mathcal{T}}$ and we create a tensor with partial information $\mathcal{T}$ by erasing, in the first half of the temporal slices of $\hat{\mathcal{T}}$, the elements related to Np_{nodes} chosen at random. To test the limits of our method, we also consider cases where either all the nodes lost some information or a fraction of the nodes lost all their activity. More details on the cases dealt with are provided below.

A. Approximated network

For each data set and for each data loss scenario, we perform the approximation of $\mathcal{T}$, i.e., the minimization of Eq. (1), with a number of components selected using the so-called core consistency diagnostic [48] as a guide (see Sec. VII for details). We perform 20 decompositions in each case, varying the initial conditions, and we use the one with the highest core consistency value. Moreover, for each value of p_{nodes} we repeat the procedure on ten different sets of nodes with missing information chosen at random, in order to evaluate the variations in the performance of our method.

TABLE I. Range of values of the Pearson coefficient computed by comparing the original and approximated node activities for each node in the different sets considered. The table also reports the median value obtained. The p values are lower than 10^{-3}.

p_{nodes}	p_{times}	Pearson's coefficient	Median value
		LSCH	
0.1	0.5	[0.53,0.96]	0.77
0.2	0.5	[0.54,0.96]	0.76
0.4	0.5	[0.53,0.96]	0.77
		HT09	
0.1	0.5	[0.38,0.83]	0.54
0.2	0.5	[0.37,0.80]	0.52
0.4	0.5	[0.37,0.73]	0.50
		SFHH	
0.1	0.5	[0.40,0.89]	0.57
0.2	0.5	[0.48,0.93]	0.57
0.4	0.5	[0.40,0.89]	0.57

From the factorization of the tensors with missing information $\mathcal{T}_{\text{HT09}}$, $\mathcal{T}_{\text{SFHH}}$, and $\mathcal{T}_{\text{LSCH}}$, we recover the first approximated version of the complete network $\mathcal{T}_{\text{app}}$ for each data set. The first evaluation of the results provided by the decomposition is obtained by computing the Pearson coefficient between the complete and approximated temporal activities of the nodes for which only partial information is available in $\mathcal{T}$. The correlation is measured only on the part of the activity that is missing in the partial data. We report the correlation coefficients found in Table I and we show in Appendix A a representative example of the temporal activities of these nodes with partial information in the complete, partial, and approximated networks. The correlation coefficients shown are measured on ten different sets of nodes with missing information for each data set. The ranges of Pearson's coefficient indicate positive moderate to high correlation between the complete and approximated node activity, indicating a good recovery of the node temporal activities in $\mathcal{T}_{\text{app}}$. Interestingly, the results depend only weakly on the fraction of nodes affected by information loss. This is due to the fact that these nodes are here chosen at random in the resampling procedure. They are thus distributed among all the latent structures present in the data so that each of these structures will also retain a number of nodes with no missing data, making it still detectable by the NTF step and allowing us to access the whole activity timeline of each structure. By definition of the latent structures as subnetworks with correlated link activities, this in turn implies that the NTF yields a good approximation of the activity timeline of the nodes affected by missing data. Obviously, a scenario in which whole latent structures would be missing from the data would yield a worse recovery of the activity timelines of the nodes with incomplete data. In addition, we note that, by construction, the approximation through the factorization relies on the

existence of correlated activity patterns and consequently performs better in the presence of strong such patterns. This explains why the temporal activities are better recovered (larger Pearson correlation coefficients) in the LSCH case than in HT09 and SFHH. Indeed, in schools the schedule is quite constrained and all the students of a given class have highly correlated activity timelines, leading to stronger correlated activity patterns than in conferences, during which attendees are free to move and interact with different people at different times.

B. Surrogate network

As illustrated in the preceding section, the approximation step achieves good results to recover single-node temporal activities by using NTF handling missing information. Despite these correlations, however, the direct use of the approximated network $\mathcal{T}_{\text{app}}$ in simulations of spreading processes results in a strong underestimation of the process outcomes (see the Appendix B). We thus apply the procedure described above and build a surrogate temporal network $\mathcal{T}_{\text{surrogate}}$ relevant to estimate the outcome of spreading processes. For the JNTF, for each link involving nodes with partial information, we moreover assign a weight extracted from the distribution of weights measured on the links with no missing information and, if needed, we remove a part of the link's activity present in $\mathcal{T}_{\text{surrogate}}$ but not in $\mathcal{T}$ until we match this weight.

C. Estimate of the outcome of spreading processes

To evaluate the performance of the method in estimating the outcome of spreading processes on the network, we simulate susceptible-infected-recovered (SIR) processes (see details in Sec. VII E) over three temporal networks for each data set: the complete, the partial, and the surrogate network. In each case we run multiple simulations for each set of values (β,μ) of the infection and recovery probabilities per unit time. We focus on the couples of probabilities (β,μ) that satisfy the following criterion. The spreading has to be finished within the time span of the data set and the epidemic size, defined as the final fraction of recovered individuals, has to be greater than 20% and lower than 80% (the selection is based on the median of the epidemic size). These conditions ensure that we avoid the selection of parameters such that the simulations either never reach a significant epidemic size and/or are too slow with respect to the total time span of the network. The limit for the final number of recovered individuals to 80% of the entire population prevents the selection of parameters leading to too fast spreading that are far from realistic conditions.

For each selected pair of parameter values β and μ, we compute the distribution of the epidemic sizes in the three cases (here called complete, partial, and surrogate). As we simulate a loss of data by considering ten different sets of randomly chosen nodes with partial information, we report the median distribution of the epidemic size on the simulations over these ten cases as well as the 25th and 75th percentiles. In addition, we provide a quantitative evaluation of the results by measuring the Jensen-Shannon divergence between the distribution of epidemic sizes obtained on the original network and on the surrogate cases. The Jensen-Shannon divergence quantifies the difference between two distributions P_1 and P_2 of epidemic sizes σ in the following way:

$$D^s_{\text{JS}}(P_1\|P_2) = \frac{1}{2}\sum_{\sigma} P_1(\sigma)\ln\frac{P_1(\sigma)}{P_2(\sigma)} + \frac{1}{2}\sum_{i} P_2(\sigma)\ln\frac{P_2(\sigma)}{P_1(\sigma)}.$$

We first consider the case in which the only available information is the partial temporal network $\mathcal{T}$ and then a case in which an additional source of information is available, namely, the location of individuals over time.

1. Using only the partial temporal network

In Fig. 2 we report the epidemic size distributions obtained for the LSCH data set, for two couples of selected spreading parameters, $\beta = 0.3$ and $\mu = 0.3$ [Figs. 2(a)–2(c)] and $\beta = 0.15$ and $\mu = 0.25$ [Figs. 2(d)–2(f)], and for different fractions of nodes p_{nodes} with partial information. The figure shows that the distributions obtained by simply simulating the SIR process on the network with partial information strongly underestimates the epidemic size and this underestimation becomes stronger as p_{nodes} increases. The distributions obtained using the baseline network is in good agreement with the ones obtained with the full original network at low p_{nodes}, but the estimation becomes rapidly rather bad as p_{nodes} increases. On the other hand, using the surrogate network yields consistently much better results. This is also corroborated by the differences between the distributions measured with the Jensen-Shannon divergences and reported in Table II.

TABLE II. Jensen-Shannon divergences measured between the distributions of epidemic sizes obtained on each network (partial, surrogate, and baseline) and on the original network.

p_{nodes}	p_{times}	Partial	Baseline	Recovered
		LSCH ($\beta = 0.3$, $\mu = 0.3$)		
0.1	0.5	0.17	0.03	0.01
0.2	0.5	0.35	0.03	0.01
0.4	0.5	0.39	0.04	0.06
		LSCH ($\beta = 0.15$, $\mu = 0.25$)		
0.1	0.5	0.17	0.01	0.01
0.2	0.5	0.34	0.01	0.01
0.4	0.5	0.34	0.08	0.02
		HT09 ($\beta = 0.6$, $\mu = 0.1$)		
0.1	0.5	0.12	0.02	0.02
0.2	0.5	0.52	0.03	0.02
0.4	0.5	0.6	0.15	0.08
		HT09 ($\beta = 0.25$, $\mu = 0.05$)		
0.1	0.5	0.15	0.01	0.05
0.2	0.5	0.35	0.08	0.02
0.4	0.5	0.51	0.32	0.06
		SFHH ($\beta = 0.3$, $\mu = 0.1$)		
0.1	0.5	0.47	0.08	0.06
0.2	0.5	0.6	0.33	0.14
0.4	0.5	0.58	0.48	0.39
		SFHH ($\beta = 0.25$, $\mu = 0.08$)		
0.1	0.5	0.32	0.07	0.06
0.2	0.5	0.48	0.21	0.13
0.4	0.5	0.49	0.34	0.24

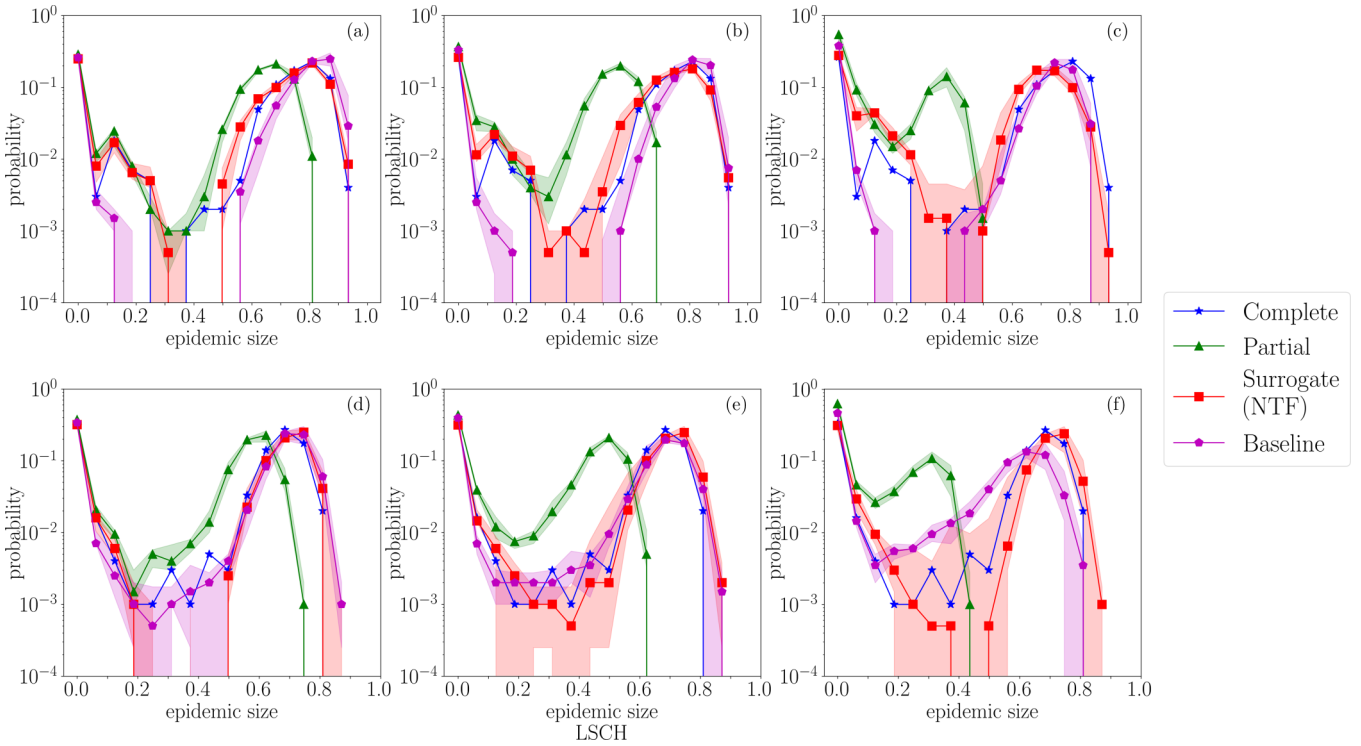

FIG. 2. Distributions of epidemic sizes computed in the complete, partial, baseline, and surrogate cases for the LSCH dataset. Each panel corresponds to $p_{\text{times}} = 0.5$, one value of the fraction p_{nodes} of nodes with partial information [panels (a),(d) $= 0.1$; (b),(e) $= 0.2$; (c),(f) $= 0.4$] and one couple of spreading parameters: (a)–(c) $\beta = 0.3$ and $\mu = 0.3$; (d)–(f) $\beta = 0.15$ and $\mu = 0.25$. For the partial, baseline, and surrogate cases, the symbols and lines show the median distribution of the epidemic size computed from the results relative to the 10 different sets of nodes with partial information, while the shaded area is delimited by the 25th and 75th percentiles.

We obtain similar results for the HT09 and SFHH data sets, for which we report the results of numerical simulations of the SIR process for various values of the spreading parameters (β,μ) and of p_{nodes} in Figs. 3 and 4. In all cases, using only the partial temporal network in the numerical simulations of the SIR process leads to a clear underestimation of the epidemic sizes and this underestimation becomes worse as p_{nodes} increases. Using the baseline networks yields better results at small p_{nodes} and the simulations performed on the surrogate network are systematically much closer to the ones based on the complete information, leading thus to a strong improvement in the prediction of the epidemic risk. For the SFHH case, however, we observe that the agreement becomes worse when p_{nodes} increases even for the surrogate network. This is due to the fact that during large conferences attendees tend to follow the schedule less rigorously than in small ones and move around and engage in contacts more freely and in a more random way, thus leading to less correlated activity patterns. In such a case, auxiliary data could be useful, as in the case we describe in the next section. Table II gives a quantitative description of the results.

2. *Using both the partial network and an additional proxy*

By construction, the method based on the NTF cannot handle extreme cases such as missing information for all nodes at the same time or activity fully missing for some nodes (as considered in [20]). To address this limitation, we propose an extension of our method that makes it possible to take advantage of additional information that can be used as a proxy for human proximity. To this aim, we consider the JNTF method described in Sec. IV B. We test this extended method on the LSCH data set, for which we have access to the approximate position of individuals in time. There are indeed 15 locations in the school: ten classes, the cafeteria, the playground, two staircases, and a control room. The resulting bipartite temporal network relating individuals and locations is composed of $N = 241$ nodes representing individuals, 15 nodes representing locations, and 130 temporal snapshots; the tensor representing this additional information has dimensions $I = 241$, $J = 15$, and $K = 130$. We note that due to the temporal resolution selected, a node might appear in several locations in the same snapshot.

To compare the methods of construction of surrogate data based on the JNTF and on the NTF decompositions, we simulated a loss of data on the contact network for a fraction of nodes $p_{\text{nodes}} = 0.2$ and for several fractions of the time span $p_{\text{times}} = [0.6, 0.8, 1]$ selected consecutively on the temporal activity of the nodes. Here the set of nodes with partial activity is the same for all values of p_{times}, so we can compare the outcomes of the SIR process and the impact of incrementally removing larger fractions of the temporal activity of the same nodes. We build surrogate data using the methods based on the NTF and on the JNTF decompositions, performed, respectively, on the partial contact networks and on the joint partial contact and location networks. For the JNTF, we impose a coupling on the first and third dimensions (i.e., we impose the equality of the matrices obtained in the decompositions

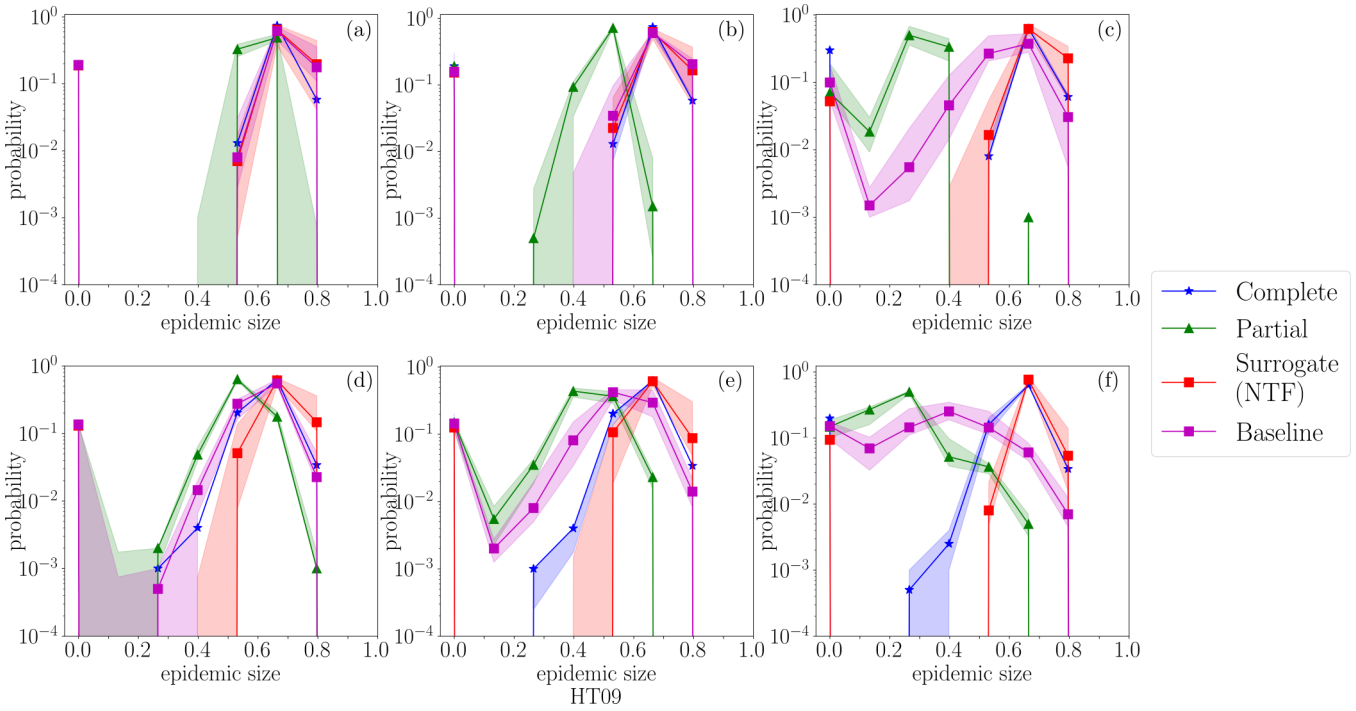

FIG. 3. Distributions of epidemic sizes computed in the complete, partial, baseline, and surrogate cases for the HT09 dataset. Each panel corresponds to $p_{\text{times}} = 0.5$, one value of the fraction p_{nodes} of nodes with partial information [panels (a),(d) = 0.1; (b),(e) = 0.2; (c),(f) = 0.4] and one couple of spreading parameters: (a)–(c) $\beta = 0.60$ and $\mu = 0.10$; (d)–(f) $\beta = 0.25$ and $\mu = 0.05$. The symbols and lines show the median distribution of the epidemic size computed from the results relative to the 10 different sets of nodes with partial information, while the shaded area is delimited by the 25th and 75th percentiles.

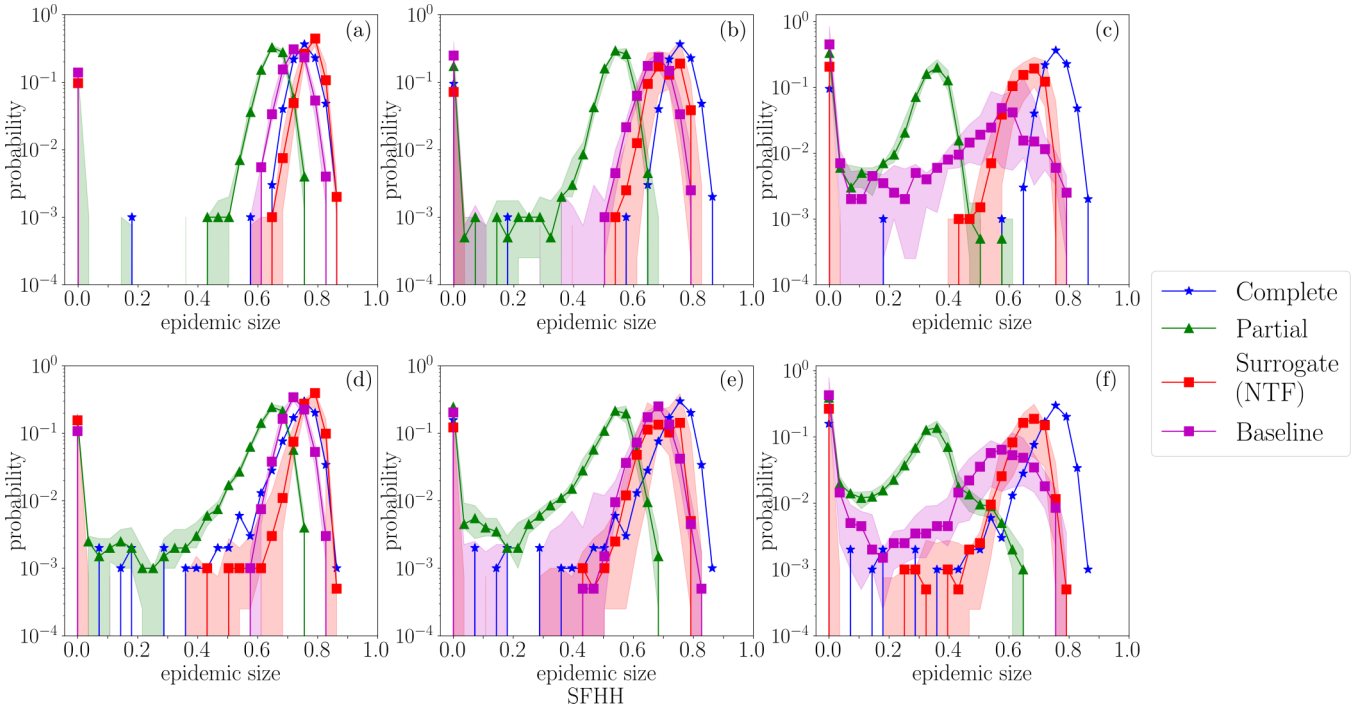

FIG. 4. Distributions of epidemic sizes computed in the complete, partial, baseline, and surrogate cases for the SFHH dataset. Each panel corresponds to $p_{\text{times}} = 0.5$, one value of the fraction p_{nodes} of nodes with partial information [panels (a),(d) = 0.1; (b),(e) = 0.2; (c),(f) = 0.4] and one couple of spreading parameters: (a)–(c) $\beta = 0.3$ and $\mu = 0.10$; (d)–(f) $\beta = 0.25$ and $\mu = 0.08$. For the partial, baseline and surrogate cases, the symbols and lines show the median distribution of the epidemic size computed from the results relative to the 10 different sets of nodes with partial information, while the shaded area is delimited by the 25th and 75th percentiles.

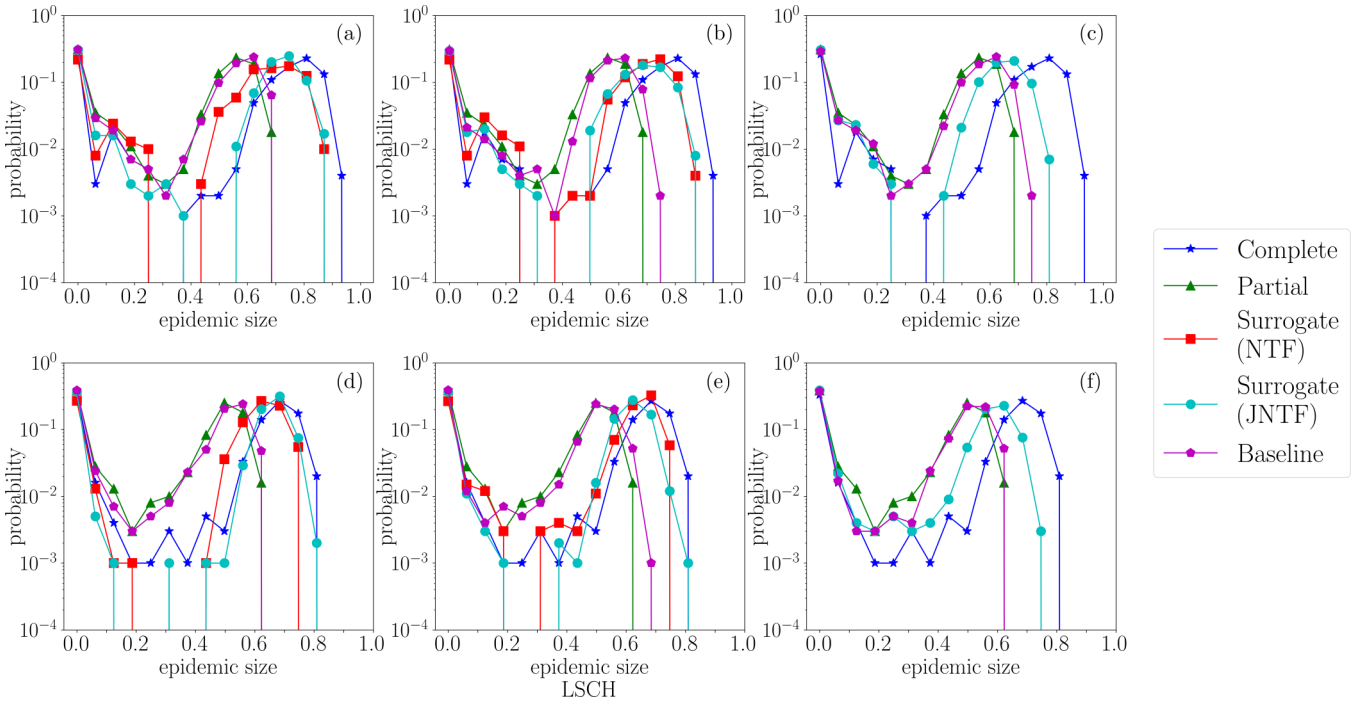

FIG. 5. Distributions of epidemic sizes computed in the complete, partial, baseline, and surrogate cases both using the NTF with only the partial network and also using the JNTF with auxiliary data for the LSCH data set. Each panel corresponds to a fixed fraction of nodes $p_{\text{nodes}} = 0.2$ with increasing loss of information for nodes $p_{\text{times}} = 0.6, 0.8, 1$. The JNTF is computed on the two tensors representing the temporal social network of the LSCH data set and the related temporal location network. Here we report the epidemic size distributions of the complete, partial, and surrogate networks for SIR processes with two different couples of infection and recovery probabilities: (a)–(c) $\beta = 0.3$ and $\mu = 0.3$ and (d)–(f) $\beta = 0.15$ and $\mu = 0.25$. In (c) and (f), as $p_{\text{times}} = 1$, no information can be gained on the nodes with missing data using simply NTF, so the results of the NTF-based method coincide with the ones obtained by simulating the SIR process over the network with partial information.

of the two tensors, for $\mathbf{A}$ on the one hand and for $\mathbf{C}$ on the other hand: $\mathbf{A}_1 = \mathbf{A}_2$ and $\mathbf{C}_1 = \mathbf{C}_2$), as the two networks have the same nodes related to individuals and the same snapshots in time. The choice of this coupling relies on the reasonable assumption that co-located individuals are more likely to be in contact. Note that $\mathbf{B}_1$ is used as in the NTF case to compute the membership of links to structures. On the other hand, $\mathbf{B}_2$ could be used with $\mathbf{A}_2$ to compute the membership of locations to latent structures; as we are interested here in computing a surrogate network for the contact data, we do not use this information in our method.

In Fig. 5 we display the results obtained for $p_{\text{nodes}} = 0.2$, $p_{\text{times}} = 0.6, 0.8, 1$, and SIR processes with the infection and recovery probabilities $\beta = 0.3$ and $\mu = 0.3$ [Figs. 5(a)–5(c)] and $\beta = 0.15$ and $\mu = 0.25$ [Figs. 5(d)–5(f)]. As in the previous cases, the distributions of epidemic sizes computed on the network with partial information show a clear underestimation of the number of individuals infected. For these large values of p_{times}, using the baseline network also yields a strong underestimation of the epidemic sizes. For $p_{\text{times}} = 0.6$ and $p_{\text{times}} = 0.8$, the results achieved with both procedures based on the NTF and the JNTF are in agreement and give a better estimation of the distributions of epidemic sizes obtained in the original network than using the incomplete data. Finally, we report in Figs. 5(c) and 5(f) the extreme case for which no information about 20% of nodes is present in the partial contact network ($p_{\text{times}} = 1$). In this case, as no information is present at all for the selected nodes, no correlated activity pattern concerning them can be inferred by the NTF decomposition, which thus cannot help recover any information on these nodes. Using JNTF then proves helpful and the surrogate network built using this method, which combines contact and location information, yields a distribution of epidemic size closer to the one obtained on the complete network. This shows that the external information provided by the approximated location of individuals in the school helps to infer possible correlations in the contact activity, even for nodes for which no contact activity was initially known. We note, however, that the epidemics sizes remain underestimated with respect to the complete network. All these results are confirmed with the Jensen-Shannon divergence computed between the distribution obtain on the complete network and the other distributions respectively obtained with the baseline and the two surrogate networks (see Tables III and IV).

Finally, we consider a case in which all nodes had some activity missing. We simulate a scenario in which all the nodes have lost half of their activity: For each node, we zero out its activity for half of the times taken at random among the times it was active. In that way each node has a different set of times during which its activity is erased. We apply the method based on the JNTF in such scenarios with ten random generations of the missing times for each node. The epidemic size distributions obtained are represented on Fig. 6 together with the one measured with the complete network. While the distributions measured on the partial networks strongly differ from those obtained with the complete network, the approach

TABLE III. Jensen-Shannon divergences measured between the distributions of epidemic sizes obtained on each network [partial, surrogate (both NTF, when this is possible, and JNTF), baseline] and on the original network.

p_{nodes}	p_{times}	Partial	Baseline	Recovered (NTF)	Recovered (JNTF)
			LSCH ($\beta = 0.3$, $\mu = 0.3$)		
0.2	0.6	0.36	0.32	0.08	0.06
0.2	0.8	0.36	0.32	0.08	0.09
0.2	1	0.36	0.31		0.19
			LSCH ($\beta = 0.15$, $\mu = 0.25$)		
0.2	0.6	0.33	0.3	0.06	0.02
0.2	0.8	0.33	0.3	0.04	0.09
0.2	1	0.33	0.31		0.15

based on the JNTF yields a much better estimation of the real epidemic size distribution.

VI. DISCUSSION

In this work, we have proposed a versatile approach to face the problem of estimating the outcome of spreading processes on temporal human proximity networks built from incomplete information. Our method leverages the existent correlations in the observed activity of the nodes to recover the contact properties of nodes whose activity is partially missing, without depending on the availability of metadata describing the nodes. To this aim, we rely on tensor decomposition techniques able to extract the mesoscale properties of temporal networks. In practice, the methodology we put forward follows two main steps: (i) the extraction of structures from the partial network through tensor decomposition and (ii) the construction of a surrogate network that is used in numerical simulations to estimate the outcome of spreading processes.

In the first step we use NTF handling missing values to decompose the partial networks into a sum of components, each describing a latent structure. In this step, we take into account

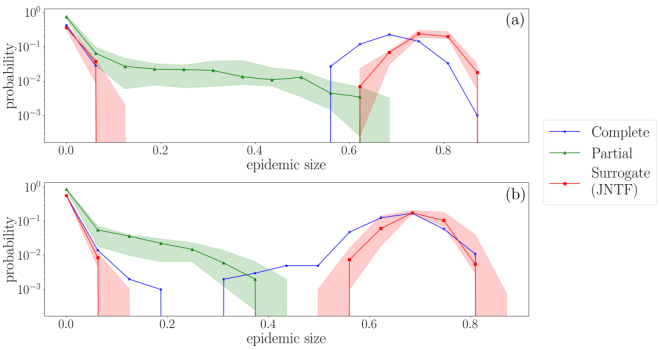

FIG. 6. Results achieved by the surrogate construction method using the JNTF. Each panel corresponds to the case in which $p_{\text{nodes}} = 1$ and $p_{\text{times}} = 0.5$. The JNTF is computed on the two tensors representing the temporal human proximity network of the LSCH data set and the related temporal location network. Here we report the results for two different couples of infection and recovery probabilities: (a) $\beta = 0.3$ and $\mu = 0.3$ and (b) $\beta = 0.15$ and $\mu = 0.25$. The symbols and lines show the median distributions of epidemic sizes computed over the ten random generations of the missing times, while the shaded area is delimited by the 25th and 75th percentiles.

the fact that the tensor describing the temporal contact network is based on incomplete information and the decomposition is thus performed using only the part of the tensor composed of known contacts or known noncontacts. This leads to a good approximation of the temporal activity of the nodes with partial information. In the second step, we determine which links belong to each latent structure and use the binarized activity timeline of each structure to fill in the unknown part of the timeline of each link with missing information. We obtain in this way a surrogate network comparable to the empirical one when used in data-driven models of epidemic spread.

We have indeed tested our method on three different data sets describing face-to-face contacts between individuals in different contexts (two conferences and a primary school), represented as temporal human proximity networks, on which we simulated a loss of data by resampling experiments, namely, a loss of information concerning a fraction of the individuals, each for a fraction of the total timeline (either at the same time or at different moments for the different individuals). Due to the data loss, epidemic sizes are strongly underestimated when we simulate the process on the networks with incomplete information. A simple baseline method based only on the average activity of nodes yields good results when only a few nodes are affected by missing information, but its performance decreases strongly as p_{nodes} increases. On the other hand, our method is able to correctly estimate the outcome of a spreading process even when half of the activity is missing for 40% of the nodes.

The performance of the method based on the NTF, however, decreases when the amount of missing information drastically increases. In particular, it is not applicable if no information at all is available for some of the nodes. To deal with this issue, we have proposed an adaptation of the method based on the joint factorization of multiple tensors. The JNTF is indeed a natural extension that allows us to integrate information encoded in multiple networks. It is particularly adequate if the information available concerns, on the one hand, the contacts of individuals and, on the other hand, their (approximate) location, encoded in a bipartite temporal network in which a link is drawn between a node representing an individual and a node representing a location when the individual is detected in that location. The joint factorization of the tensors representing the partial temporal network of contacts and the temporal network encoding positions, constrained to extract latent structures with the same nodes and the same activity timelines in both tensors, can then help to recover the missing information about the contacts. However, the JNTF might overestimate the activity of the links with missing activity and leads to a distribution of weights (number of contacts per pair of individuals) that is more homogeneous than in the original case. As the heterogeneity properties of contact networks are well known to play a crucial role in determining the outcome of spreading processes [49–55], we adjust the weight distribution of the links for which only partial information was available; this step is made possible by the robustness under sampling of the contact network weight distribution. We can thus rely on the weight distribution measured on those links for which no information is missing in the incomplete network to extract at random values and assign them to the links with missing information. We have tested this alternative method in a data set for which both contacts and approximated positions of nodes in

time are available and shown that it yields results similar to the NTF when both methods can be applied. When information about some nodes is completely lost, i.e., when the method based on NTF alone cannot be used, using JNTF allows us to recover part of the missing information and yields a distribution of epidemic sizes closer to the original than when using the partial network in the simulations, even when as much as 50% of the contact activity of each node has been lost in the data.

The methodology we have presented manages to cope with missing information in various contexts. Importantly, and contrarily to other methods, it does not rely on the availability of metadata or any knowledge on the structure of the population into groups (such as classes in a school) as in [20]: The NTF or JNTF decompositions are indeed able to extract the effective structure (both in groups and temporal) of the temporal network of interactions and to assign each individual to one or multiple latent structures, each with its activity timeline.

Some limitations of our method stem from the tensor decomposition itself. For instance, if a whole group of correlated links or nodes is missing in the data, it will not be uncovered by the NTF. In such an unfavorable case, the JNTF method can compensate for the lack of correlated activity in the incomplete contact network by relying on auxiliary data when available. Moreover, when the temporal network of contacts lacks structure or when the nodes with partial information behave in a random way, i.e., do not exhibit any activity correlation with other nodes, neither decomposition (NTF or JNTF) might be able to clearly assign them to any latent structure and thus to determine their activity timeline. A way to cope with this issue might be to attribute some random activity (based, for instance, on the average behavior) to those nodes with missing information that are not found in any latent structure. Given the relatively good results obtained when using the baseline network, built by assigning a random average activity to nodes with missing data, this could be a meaningful starting point to combine with our approach in order to better tackle the problem of missing data for nodes whose activity turns out to be completely uncorrelated with other nodes.

Another limitation, which can be easily overcome though, is that for the JNTF we rely on the distribution of link weights measured in the partial network. Indeed, if the remaining information is not representative enough (too much information missing), the heterogeneity properties of the surrogate network might be affected as the weights assigned to the links with missing information are taken among those measured. We could however take advantage of the known robustness of the weight distributions in different contexts [15,56] to use publicly available weight distributions collected in other contexts.

Finally, a natural direction for future research is the extension of our technique to infer mesoscale properties of human proximity networks even when no direct information on contacts is available, but only proxies such as approximate locations of individuals have been collected.

VII. METHODOLOGY

Here we describe some technical details regarding the different steps of our procedure.

A. The NTF computation

To build the surrogate network, we need to apply the non-negative tensor factorization to the masked tensor introduced in this paper. We consider an algorithm based on an alternating large-scale non-negativity-constrained least-squares framework using the Karush-Kuhn-Tucker (KKT) optimality conditions in a block principal pivoting (BPP) framework [43]. To solve the problem through the BPP algorithm we rewrite the minimization problem as a multiple-right-hand-side problem of the form

$$\min \|\mathbf{VX} - \mathbf{W}\|_F^2. \tag{3}$$

This is done using matricization, which transforms the problem in three minimization subproblems [32]. Then we follow the update rules given by [43]. This allows us to perform the original non-negative tensor factorization. As we mask a part of the tensor, however, we need to adapt the method. In practice, at each time step of the factorization of $\mathcal{W} * \mathcal{T}$, we replaced the masked elements by the corresponding elements in the tensor $[[\boldsymbol{\lambda}_t; \mathbf{A_t}, \mathbf{B_t}, \mathbf{C_t}]]$, where $\lambda_t; \mathbf{A_t}, \mathbf{B_t}, \mathbf{C_t}$ are the latent factors at the iteration step t considered.

B. The JNTF computation

To integrate data from multiple sources we use the joint non-negative tensor factorization, described in Sec. VII B. To this aim, we adapt the alternating large-scale non-negativity-constrained least-squares framework and in particular the way to compute the KKT optimality conditions in a BPP framework [43]. Here we illustrate the adaptation details to solve Eq. (2) in the case of two data sources that are coupled in two dimensions (case study in Sec. V), i.e., $S = 2$, $\mathbf{A} = \mathbf{A}_1 = \mathbf{A}_2$, and $\mathbf{C} = \mathbf{C}_1 = \mathbf{C}_2$, such that Eq. (2) becomes

$$\begin{aligned} \min \Bigg(& \frac{1}{2}\|\mathcal{T}_1 - [[\boldsymbol{\lambda}_1; \mathbf{A}, \mathbf{B}_1, \mathbf{C}]]\|_F^2 + \frac{\alpha_1}{2}\|\boldsymbol{\lambda}_1\|_2^2 \\ & + \frac{1}{2}\|\mathcal{T}_2 - [[\boldsymbol{\lambda}_2; \mathbf{A}, \mathbf{B}_2, \mathbf{C}]]\|_F^2 + \frac{\alpha_2}{2}\|\boldsymbol{\lambda}_2\|_2^2 \Bigg) \\ & \text{such that } \boldsymbol{\lambda}_{1,2}, \mathbf{A}, \mathbf{B}_{1,2}, \mathbf{C} \geqslant 0. \end{aligned} \tag{4}$$

Note that we added here the regularization terms $\frac{\alpha_{1,2}}{2}\|\boldsymbol{\lambda}_{1,2}\|_F^2$, which are sparsity penalties.

To solve the problem through the BPP algorithm we rewrite the minimization problem as a multiple-right-hand-side problem of the form

$$\min \|\mathbf{VX} - \mathbf{W}\|_F^2 \tag{5}$$

and solve it for each of the factor matrices (here we include $\boldsymbol{\lambda}_{1,2}$ in the factor matrices). To this aim, we start by solving the problem for the factor matrices $\mathbf{B}_1$ and $\mathbf{B}_2$ which are respectively present in the first and third terms of Eq. (4). With respect to these factor matrices, we can rewrite the equation by using the two-mode matricization [32] of $\mathcal{T}_1$ and $\mathcal{T}_2$, which leads to the approximations

$$\mathbf{T}_{1,(2)} \approx \mathbf{B}_1 \boldsymbol{\Lambda}_1 (\mathbf{C} \odot \mathbf{A})^T,$$

$$\mathbf{T}_{2,(2)} \approx \mathbf{B}_2 \boldsymbol{\Lambda}_2 (\mathbf{C} \odot \mathbf{A})^T,$$

where

$$\boldsymbol{\Lambda}_1 = \mathrm{diag}(\lambda_{1,1}, \ldots, \lambda_{1,R}),$$

$$\boldsymbol{\Lambda}_2 = \mathrm{diag}(\lambda_{2,1}, \ldots, \lambda_{2,R}).$$

We can rewrite the approximations as

$$\mathbf{T}_{1,(2)}^T \approx (\mathbf{C}\odot\mathbf{A})\mathbf{\Lambda}_1\mathbf{B}_1^T,$$
$$\mathbf{T}_{2,(2)}^T \approx (\mathbf{C}\odot\mathbf{A})\mathbf{\Lambda}_2\mathbf{B}_2^T,$$

where $\mathbf{\Lambda}_{1,2}^T = \mathbf{\Lambda}_{1,2}$ and thus

$$\mathbf{B}_1^T \approx \mathbf{\Lambda}_1^{-1}(\mathbf{C}\odot\mathbf{A})^\dagger\mathbf{T}_{1,(2)}^T,$$
$$\mathbf{B}_2^T \approx \mathbf{\Lambda}_2^{-1}(\mathbf{C}\odot\mathbf{A})^\dagger\mathbf{T}_{2,(2)}^T.$$

By using the property of the Khatri-Rao product, for which

$$(\mathbf{C}\odot\mathbf{A})^\dagger = (\mathbf{C}^T\mathbf{C} * \mathbf{A}^T\mathbf{A})^\dagger(\mathbf{C}\odot\mathbf{A})^T,$$

we can write the approximation as

$$\mathbf{\Lambda}_1(\mathbf{C}^T\mathbf{C} * \mathbf{A}^T\mathbf{A})\mathbf{B}_1^T \approx (\mathbf{C}\odot\mathbf{A})^T\mathbf{T}_{1,(2)}^T,$$
$$\mathbf{\Lambda}_2(\mathbf{C}^T\mathbf{C} * \mathbf{A}^T\mathbf{A})\mathbf{B}_2^T \approx (\mathbf{C}\odot\mathbf{A})^T\mathbf{T}_{2,(2)}^T.$$

The related subproblems of Eq. (4) for $\mathbf{B}_1$ and $\mathbf{B}_2$ are then reduced to

$$\min_{\mathbf{B}_1} \tfrac{1}{2}\|\mathbf{\Lambda}_1(\mathbf{C}^T\mathbf{C} * \mathbf{A}^T\mathbf{A})\mathbf{B}_1^T - (\mathbf{C}\odot\mathbf{A})^T\mathbf{T}_{1,(2)}^T\|_F^2,$$
$$\min_{\mathbf{B}_2} \tfrac{1}{2}\|\mathbf{\Lambda}_2(\mathbf{C}^T\mathbf{C} * \mathbf{A}^T\mathbf{A})\mathbf{B}_2^T - (\mathbf{C}\odot\mathbf{A})^T\mathbf{T}_{2,(2)}^T\|_F^2.$$

Finally, the subproblems can be respectively written in the form of Eq. (5) by assigning

$$\mathbf{V} = \mathbf{\Lambda}_1(\mathbf{C}^T\mathbf{C} * \mathbf{A}^T\mathbf{A}), \quad \mathbf{X} = \mathbf{B}_1^T,$$
$$\mathbf{W} = \mathbf{\Lambda}_2(\mathbf{C}\odot\mathbf{A})^T\mathbf{T}_{1,(2)};$$
$$\mathbf{V} = (\mathbf{C}^T\mathbf{C} * \mathbf{A}^T\mathbf{A}), \quad \mathbf{X} = \mathbf{B}_2^T,$$
$$\mathbf{W} = (\mathbf{C}\odot\mathbf{A})^T\mathbf{X}_{2,(2)}^T.$$

The solution to the subproblems is now straightforward, as illustrated in [43].

We now need to rewrite the subproblems for the factors $\mathbf{A}$ and $\mathbf{C}$ that are present in both the first and third terms of Eq. (4). We show the procedure for $\mathbf{A}$ (which is analogous for $\mathbf{C}$). First, we write the minimization problem by using the one-mode matricization of $\mathcal{T}_1$ and $\mathcal{T}_2$ (three-mode matricization for $\mathbf{C}$):

$$\min_{\mathbf{A}} \left[\tfrac{1}{2}\|(\mathbf{C}\odot\mathbf{B}_1)\mathbf{\Lambda}_1\mathbf{A}^T - \mathbf{T}_{1,(1)}^T\|_F^2 + \tfrac{1}{2}\|(\mathbf{C}\odot\mathbf{B}_2)\mathbf{\Lambda}_2\mathbf{A}^T - \mathbf{T}_{2,(1)}^T\|_F^2\right].$$

By following the procedure shown above, we can write the problem as

$$\min_{\mathbf{A}} \left[\tfrac{1}{2}\|\mathbf{\Lambda}_1(\mathbf{C}^T\mathbf{C} * \mathbf{B}_1^T\mathbf{B}_1)\mathbf{A}^T - (\mathbf{C}\odot\mathbf{B}_1)^T\mathbf{T}_{1,(1)}^T\|_F^2 + \tfrac{1}{2}\|\mathbf{\Lambda}_2(\mathbf{C}^T\mathbf{C} * \mathbf{B}_2^T\mathbf{B}_2)\mathbf{A}^T - (\mathbf{C}\odot\mathbf{B}_2)^T\mathbf{T}_{2,(1)}^T\|_F^2\right],$$

in which we can assign

$$\mathbf{V}_1 = \mathbf{\Lambda}_1\mathbf{C}^T\mathbf{C} + \mathbf{B}_1^T\mathbf{B}_1, \quad \mathbf{W}_1 = (\mathbf{C}\odot\mathbf{B}_1)^T\mathbf{T}_{1,(1)}^T,$$
$$\mathbf{V}_2 = \mathbf{\Lambda}_2\mathbf{C}^T\mathbf{C} + \mathbf{B}_2^T\mathbf{B}_2, \quad \mathbf{W}_2 = (\mathbf{C}\odot\mathbf{B}_2)^T\mathbf{T}_{2,(1)}^T,$$
$$\mathbf{X} = \mathbf{A}^T,$$

leading to

$$f(\mathbf{X}) = \min_{\mathbf{X}}(\tfrac{1}{2}\|\mathbf{V}_1\mathbf{X} - \mathbf{W}_1\|_F^2 + \tfrac{1}{2}\|\mathbf{V}_2\mathbf{X} - \mathbf{W}_2\|_F^2). \quad (6)$$

To solve the problem in Eq. (6) we need to adapt the KKT conditions, which result in

$$\nabla f(\mathbf{X}) = \underbrace{(\mathbf{V}_1^T\mathbf{V}_1 + \mathbf{V}_2^T\mathbf{V}_2)}_{\mathbf{V}_{1,2}}\mathbf{X} - \underbrace{(\mathbf{V}_1^T\mathbf{W}_1 + \mathbf{V}_2^T\mathbf{W}_2)}_{\mathbf{W}_{1,2}},$$
$$\nabla f(\mathbf{X}) \geqslant 0, \quad \nabla f(\mathbf{X})^T\mathbf{X} = 0, \quad \mathbf{X} \geqslant 0,$$

whose solution is given by solving

$$\mathbf{X}^T\mathbf{V}_{1,2}^T - \mathbf{W}_{1,2}^T = 0.$$

Since Eq. (4) includes regularization terms, we have to adapt the KKT conditions for the minimization problem with respect to $\boldsymbol{\lambda}_1$ and $\boldsymbol{\lambda}_2$. We show the procedure for $\boldsymbol{\lambda}_1$, which is analogous for $\boldsymbol{\lambda}_2$. We consider the cost function built from the terms in which $\boldsymbol{\lambda}_1$ is involved [the first and second terms in Eq. (4)],

$$f_{\boldsymbol{\lambda}_1} = \frac{1}{2}\|\mathcal{X}_1 - [[\boldsymbol{\lambda}_1;\mathbf{A},\mathbf{B}_1,\mathbf{C}]]\|_F^2 + \frac{\alpha_1}{2}\|\boldsymbol{\lambda}_1\|_2^2,$$

and we rewrite it through the vectorization

$$f_{\boldsymbol{\lambda}_1} = \frac{1}{2}\|\mathrm{vec}(\mathbf{C}\odot\mathbf{B}\odot\mathbf{A})\boldsymbol{\lambda}_1 - \mathrm{vec}(\mathcal{X})\|_F^2 + \frac{\alpha_1}{2}\|\boldsymbol{\lambda}_1\|_2^2.$$

Minimizing the cost function $f_{\boldsymbol{\lambda}_1}$ is equivalent to minimizing $f'_{\boldsymbol{\lambda}_1}$, obtained by incorporating the regularization term as follows:

$$f'_{\boldsymbol{\lambda}_1} = \frac{1}{2}\left\|\underbrace{\begin{pmatrix}\mathrm{vec}(\mathbf{C}\odot\mathbf{B}_1\odot\mathbf{A})\\ \sqrt{\alpha_1}\end{pmatrix}}_{\mathbf{V}}\underbrace{\boldsymbol{\lambda}_1}_{\mathbf{x}} - \underbrace{\begin{pmatrix}\mathrm{vec}(\mathcal{X})\\ 0\end{pmatrix}}_{\mathbf{w}}\right\|_F^2.$$

The solution to the minimization of $f'_{\boldsymbol{\lambda}_1}$ follows from [43].

C. The NTF rank selection

The selection of the number of components R for the decomposition is guided by the core consistency diagnostic [48], which estimates to what extent the PARAFAC model $[[\boldsymbol{\lambda};\mathbf{A},\mathbf{B},\mathbf{C}]]$ with a given rank r (i.e., with a sum of r components) is appropriate to represent the data. The core consistency is a measure that has 100 as an upper bound and values above 50 are usually considered acceptable. Here we computed the core consistency values between the tensor with partial information and its approximation for $r \in [2,\ldots,R_{\max}]$, for five realizations of the optimization procedure starting from different initial conditions for $\mathbf{A}$, $\mathbf{B}$, and $\mathbf{C}$ for each value of r. We select a rank $R = R_{cc} - 1$, where R_{cc} is the smallest rank for which the core consistency value for each of the five realizations is lower than 85. This threshold is selected to ensure an approximation as faithful as possible to the original tensor. For the JNTF case, we use as a hint for the number of components to be selected the one obtained with the NTF on the temporal contact network with partial information. It is worth noting that, when the amount of available information in the data decreases, the value of R determined by the core consistency can vary. This is due to the fact that by erasing an increasing percentage of node activities, the correlated activity patterns are increasingly perturbed and might be destroyed; a smaller number of components (subnetworks composed by links having a correlated activity in time) is then detected.

D. The Otsu method

The Otsu method [44] is commonly used in image processing to recover the different levels of gray in pixels. The same idea can be used on the temporal activities of the components by thinking of them as images composed of K pixels of different values (level of activation). In particular, the method for two-dimensional functions assumes that the function given as an input contains values that follow a bimodal distribution. The method computes then the optimal threshold, defined by minimizing the intraclass variance and by maximizing the interclass variance. The resulting optimal threshold can be used to divide the values of the function into two groups.

Here we use the threshold given by the Otsu method to define whenever a component is active or not. This is done by applying the Otsu method on the temporal activity of each component. The values above the threshold correspond to the temporal activation of the component, while the values below the threshold correspond to the times in which the component is inactive.

E. The SIR processes

To simulate how an infectious disease propagates in a population and thus how the related dynamical process spreads over the network, we run a SIR process over the network. The SIR model assumes that each individual in the population can

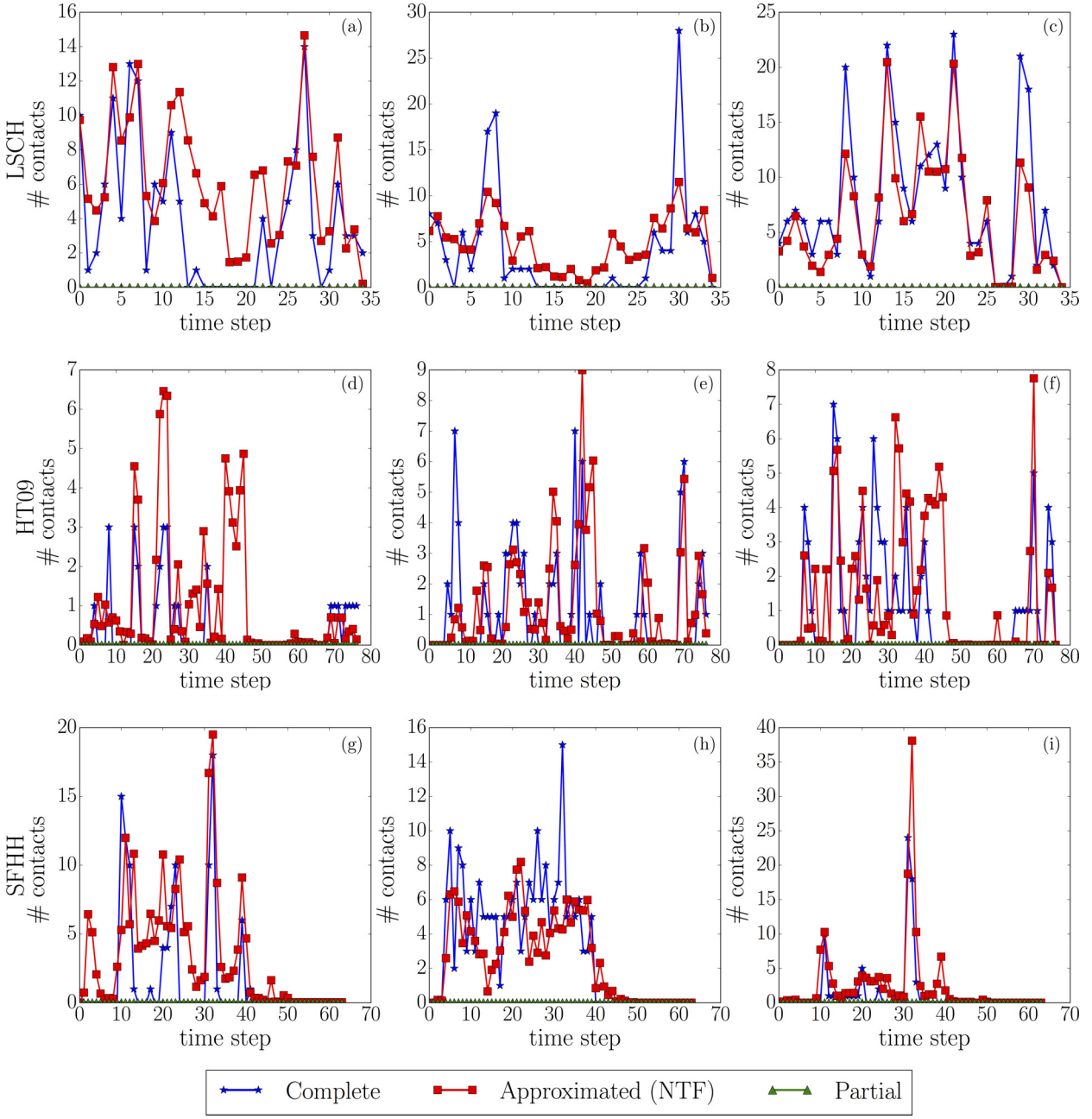

FIG. 7. Temporal activity for examples of the total number of contacts in time of a node for each data set and case study: (a)–(c) LSCH, (d)–(f) HT09, and (g)–(i) SFHH. We report the temporal activity in the complete network, in the one with partial information, and in the one obtained approximated by the NTF method. We show here the temporal activity only for the time steps in which we lost the information about the contacts of the node, corresponding in our scenario to the first half of the timeline.

be in one of three states: susceptible, infectious, or recovered. The propagation of the disease starts from a single individual, who is in the infectious state (the seed of the process), while all others are initially susceptible. At each time step, each susceptible individual in contact with an infectious one has a probability β to become infectious. Each infectious individual recovers with probability μ per time step. For simulation purposes, the first node to be infected is chosen among the nodes that are not in the set of those with partial information. For each network and each set of values of the parameters β and μ, we run 1000 simulations of the spreading process. In particular, the parameter space we consider *a priori* is given by all the couples of probabilities (β,μ), with β,μ in the range [0.001,1]. We select the suitable couples of parameters according to the conditions described in Sec. V C and we show the results corresponding to two representative examples of parameter couples for each data set. Results related to other couples of parameters are similar to the ones shown in the figures.

ACKNOWLEDGMENT

The authors acknowledge support from the Lagrange Project of the ISI Foundation funded by the CRT Foundation.

APPENDIX A: APPROXIMATED NETWORK

By decomposing a network affected by missing information through the NTF we are able to reconstruct some of the structural and temporal properties observed in the complete network. One of these characteristics is the overall contact activity of the nodes. In particular, given a node whose information is partially missing, we are able to reconstruct its overall activity in time, i.e., the number of contacts related to that node at each time. We have indeed shown that the activities in the complete network and in the approximated one are significantly correlated and have high Pearson correlation coefficients. In Fig. 7 we report a representative example for the overall temporal activity of a single node in the complete, partial, and approximated network for each data set: LSCH, HT09, and SFHH. As we consider a scenario in which the missing activity concerns, in each case, the first half of the timeline, the activity of the nodes with missing information is 0 in the partial data, while it shows nontrivial patterns in the complete network, which are partially recovered in the approximated network.

APPENDIX B: THE SIR PROCESS ON THE APPROXIMATED NETWORK

We have shown that by approximating a network with partial information via NTF we are able to recover some of its properties, such as the overall temporal activity of the nodes. However, as discussed in the main text, this information is not enough to obtain outcomes of a spreading process close to the ones obtained on the complete network, because the approximation tends to make the distributions of weights more homogeneous than the empirical one. We illustrate this fact by showing in Fig. 8 the distributions of epidemic sizes of a SIR process simulated on the complete, partial, and approximated $\mathcal{T}_{\text{app}}$ networks for the LSCH data set, for different spreading parameters $\beta = 0.30$ and $\mu = 0.30$ [Figs. 8(a)–8(c)] and $\beta = 0.15$ and $\mu = 0.25$ [Figs. 8(d)–8(f)] and different fractions of

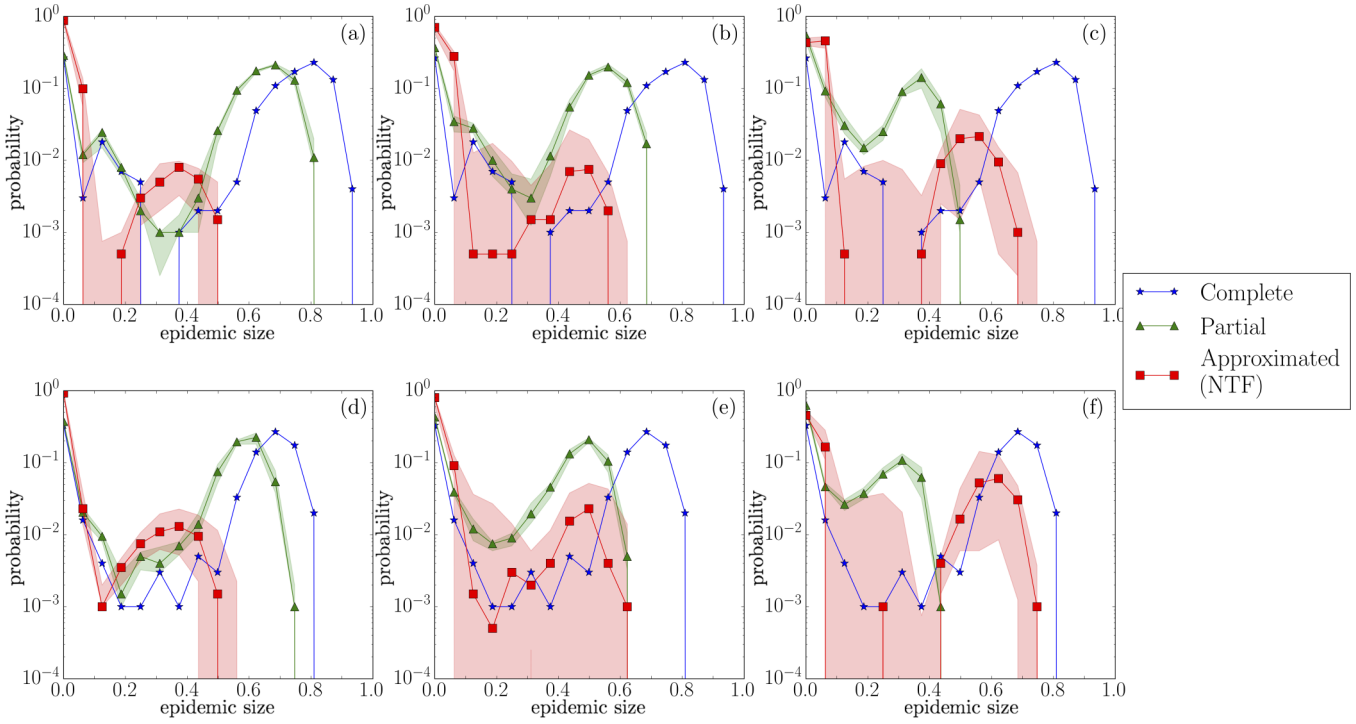

FIG. 8. Outcome of an SIR process: over the complete network, the partial network and the network approximated through the NTF for the LSCH dataset. Each panel corresponds to one fraction p_{nodes} of nodes with partial information ($a,d = 0.1$, $b,e = 0.2$, and $c,f = 0.4$) and one couple of spreading parameters: (a)–(c) $\beta = 0.30$ and $\mu = 0.30$; (d)–(f) $\beta = 0.15$ and $\mu = 0.25$. The lines show the median distribution of the epidemic size computed from the results relative to 10 different sets of nodes with missing information, while the shaded area is delimited by the 25-th and 75-th percentiles.

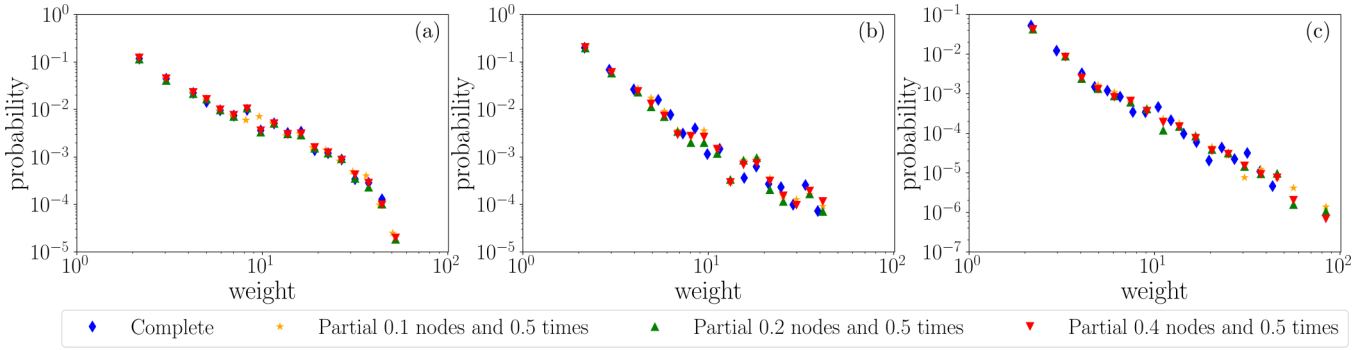

FIG. 9. Weight distributions in the (a) LSCH, (b) HT09, and (c) SFHH data sets for the complete network and the partial network with increasing percentage of nodes with missing information: $p_{\text{nodes}} \in [0.1,0.2,0.4]$ and $p_{\text{times}} = 0.5$.

nodes missing (0.1, 0.2, and 0.4) and fraction of times with missing data $p_{\text{times}} = 0.5$.

Here we report the results only for the LSCH data set, as it is the one characterized by highly correlated activity patterns and thus the one for which the approximations obtained through the NTF are closer to the complete case. In the primary school indeed, students' activity is determined by the school schedule and they are divided in classes. This makes their activity highly correlated as their contacts are more homogeneous during the daily class activities. However, as we can see from Fig. 8, even if the approximated network is close to the complete one in terms of the overall activity patterns displayed, this is not the case for the outcome of the spreading process. The epidemic sizes in the case of the partial and approximated networks are far from the complete network case. This strong underestimation is due to the fact that when we are approximating the network via NTF, the method tends to approximate the activity patterns of the nodes in the same component as fully correlated, thus connecting the nodes in the same component and rendering the network more homogeneous.

APPENDIX C: WEIGHT DISTRIBUTION

For the JNTF case, in order to obtain outcomes of SIR processes close to the case of the complete network, we have to reintroduce the heterogeneity properties into the approximated network and use the resulting surrogate network to simulate the SIR process. We reintroduce the heterogeneity properties by reassigning the weights, i.e., the total number of contacts, to the links involving nodes with partial information. In particular, we measure the weight distribution on the available part of the partial network. Then we use this distribution to pick at random a weight that will be reassigned to the link whose activity is approximated via the JNTF. We rely on such a process as the weight distribution of a network is robust to various sampling procedures, as shown in Fig. 9 for the sampling considered in this paper: For each data set, LSCH [Fig. 9(a)], HT09 [Fig. 9(b)], and SFHH [Fig. 9(c)], we compare the weight distribution of the complete network with the corresponding weight distributions computed on the links of partial networks not involving any nodes with missing information, with $p_{\text{nodes}} \in [0.1,0.2,0.4]$. By construction, these distributions do not depend on p_{times}. The results clearly show that the weight distributions measured in the partial and complete networks are consistent, meaning that, even in the case in which we miss almost half of the links in the partial network, we are able to compute a weight distribution similar to the one of the complete network.

APPENDIX D: THE SIR RESULTS FOR A GENERAL MISSING DATA SCENARIO

We consider here an additional scenario for the loss of data, in order to test the robustness of the method. We simulate on each data set a loss of data determined by a fraction of nodes $p_{\text{nodes}} \in [0.1,0.2,0.4]$ for which we zero out the activity occurring at a fraction $p_{\text{times}} = 0.5$ of the temporal snapshots, these snapshots being chosen independently at random for each node.

In such a scenario, the impact of data loss on the spreading process is lower than in the scenario considered in the main text, as Figs. 10–12 show. Indeed, the underestimation of the

TABLE IV. Jensen-Shannon divergences measured between the distributions of epidemic sizes obtained on each network (partial, surrogate, and baseline) and on the original network. See Appendix D.

p_{nodes}	p_{times}	Partial	Baseline	Recovered
		LSCH ($\beta = 0.3$, $\mu = 0.3$)		
0.1	0.5	0.029	0.08	0.01
0.2	0.5	0.09	0.14	0.02
0.4	0.5	0.17	0.19	0.04
		LSCH ($\beta = 0.15$, $\mu = 0.25$)		
0.1	0.5	0.02	0.05	0.01
0.2	0.5	0.09	0.11	0.03
0.4	0.5	0.18	0.15	0.04
		HT09 ($\beta = 0.6$, $\mu = 0.1$)		
0.1	0.5	0.02	0.03	0.08
0.2	0.5	0.10	0.05	0.17
0.4	0.5	0.27	0.09	0.25
		HT09 ($\beta = 0.25$, $\mu = 0.05$)		
0.1	0.5	0.01	0.16	0.07
0.2	0.5	0.039	0.27	0.19
0.4	0.5	0.13	0.35	0.29
		SFHH ($\beta = 0.3$, $\mu = 0.1$)		
0.1	0.5	0.05	0.04	0.11
0.2	0.5	0.16	0.05	0.11
0.4	0.5	0.21	0.12	0.15
		SFHH ($\beta = 0.25$, $\mu = 0.08$)		
0.1	0.5	0.07	0.07	0.08
0.2	0.5	0.23	0.09	0.08
0.4	0.5	0.23	0.16	0.13

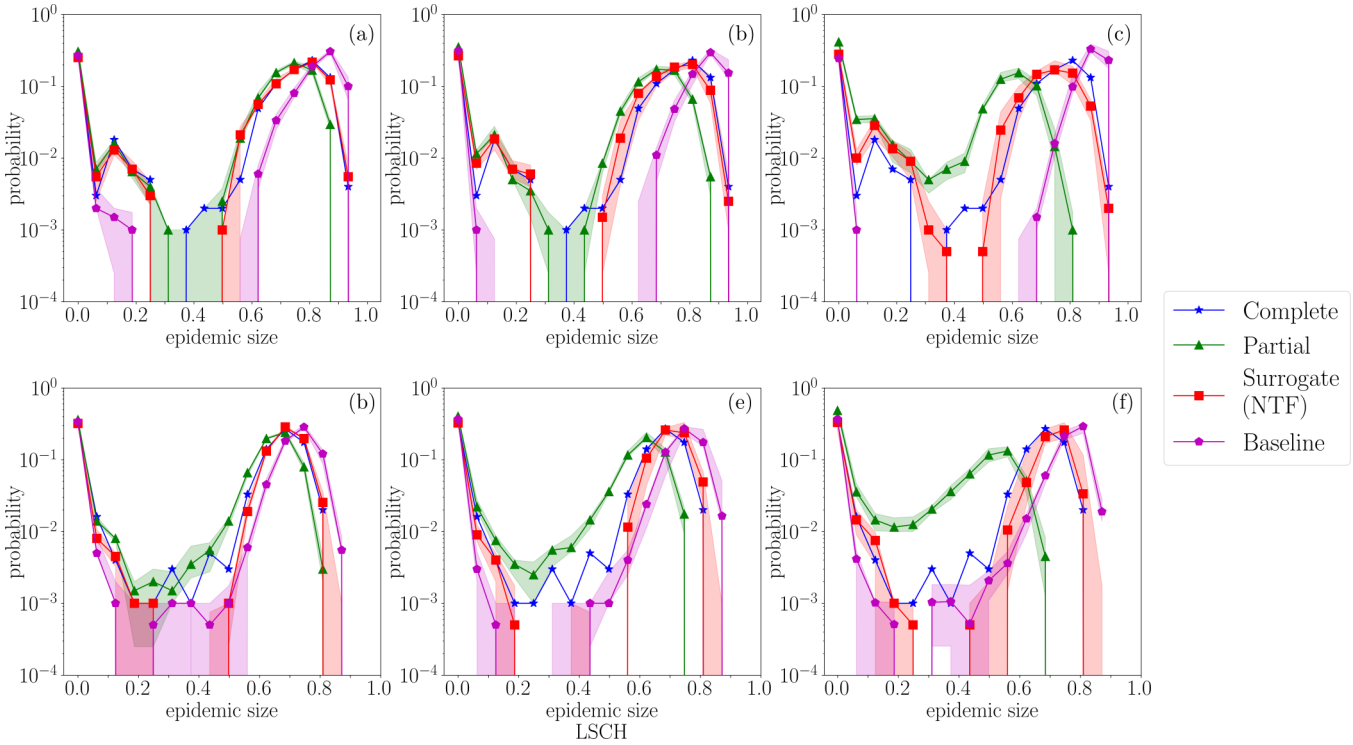

FIG. 10. Distributions of epidemic sizes computed in the complete, partial, baseline and surrogate cases for the LSCH dataset, for a scenario in which data loss occurs at random times. Each panel corresponds to one value of the fraction p_{nodes} of nodes with partial information ($a,d = 0.1$, $b,e = 0.2$, and $c,f = 0.4$) and one couple of spreading parameters: (a)–(c) $\beta = 0.3$ and $\mu = 0.3$; (d)–(f) $\beta = 0.15$ and $\mu = 0.25$. For the partial, baseline and surrogate cases, the symbols and lines show the median distribution of the epidemic size computed from the results relative to the 10 different sets of nodes with partial information, while the shaded area is delimited by the 25-th and 75-th percentiles.

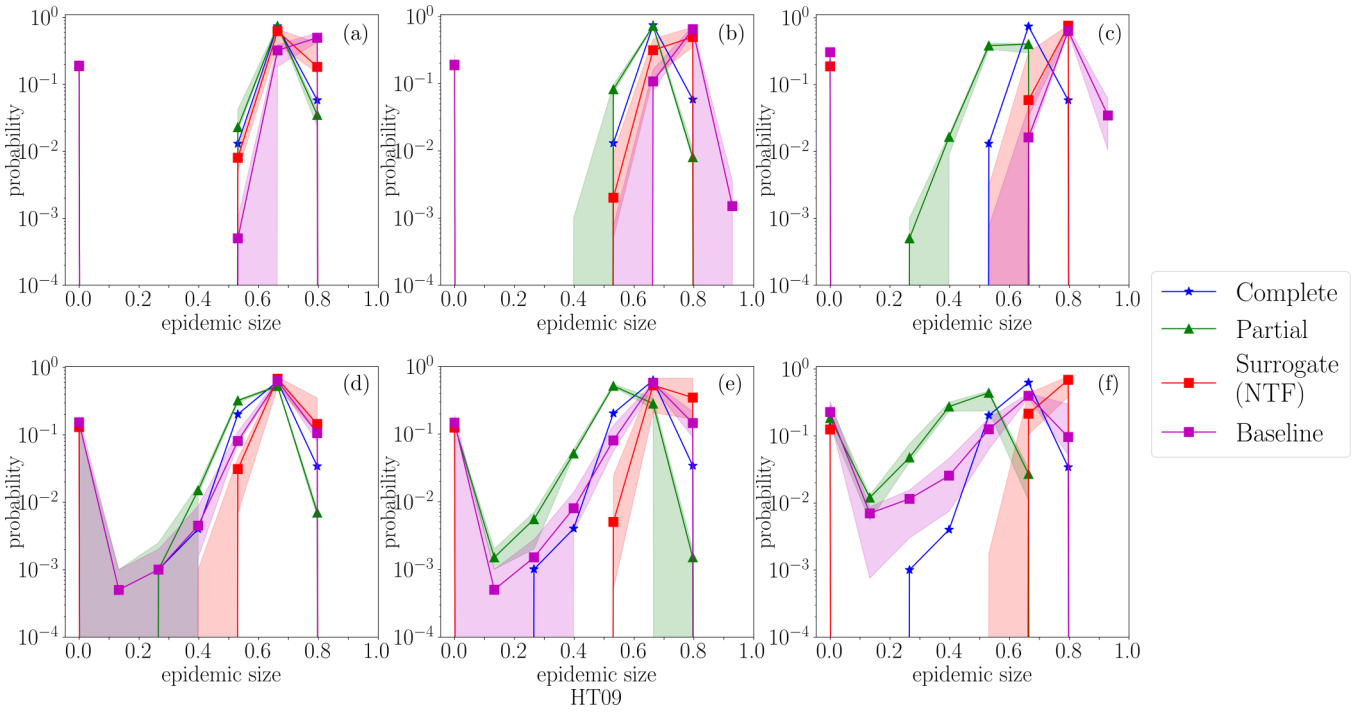

FIG. 11. Distributions of epidemic sizes computed in the complete, partial, baseline and surrogate cases for the HT09 dataset, for a scenario in which data loss occurs at random times. Each panel corresponds to one value of the fraction p_{nodes} of nodes with partial information ($a,d = 0.1$, $b,e = 0.2$, and $c,f = 0.4$) and one couple of spreading parameters: (a)–(c) $\beta = 0.60$ and $\mu = 0.10$; (d)–(f) $\beta = 0.25$ and $\mu = 0.05$. The symbols and lines show the median distribution of the epidemic size computed from the results relative to the 10 different sets of nodes with partial information, while the shaded area is delimited by the 25-th and 75-th percentiles.

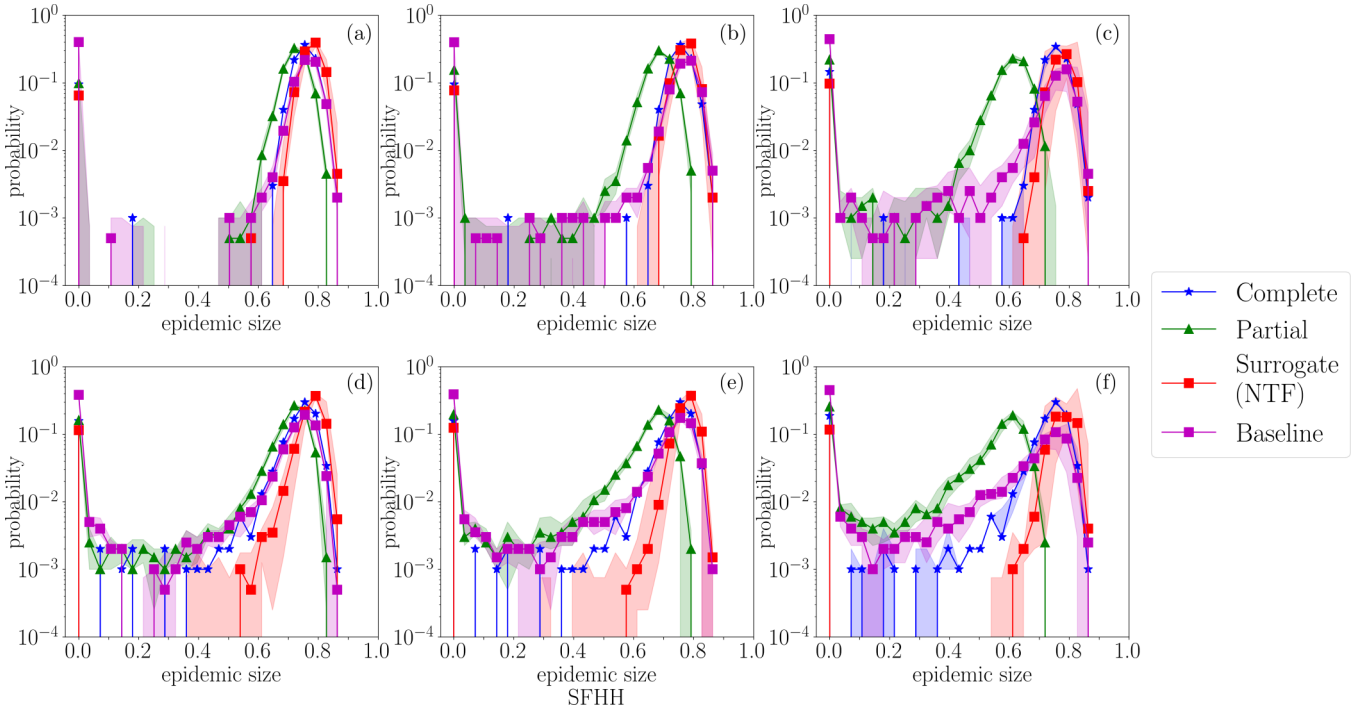

FIG. 12. Distributions of epidemic sizes computed in the complete, partial, baseline and surrogate cases for the SFHH dataset, for a scenario in which data loss occurs at random times. Each panel corresponds to one value of the fraction p_{nodes} of nodes with partial information ($a,d = 0.1$, $b,e = 0.2$, and $c,f = 0.4$) and one couple of spreading parameters: (a)–(c) $\beta = 0.3$ and $\mu = 0.10$; (d)–(f) $\beta = 0.25$ and $\mu = 0.08$. For the partial baseline, and surrogate cases, the symbols and lines show the median distribution of the epidemic size computed from the results relative to the 10 different sets of nodes with partial information, while the shaded area is delimited by the 25-th and 75-th percentiles.

epidemic sizes when using the partial network is less strong. For the LSCH case, we obtain results similar to the ones presented in the main text, with good agreement of the distributions obtained with the surrogate and complete networks. In the SFHH and HT09 data sets, baseline and surrogate networks yield results that are closer. This can be explained by (i) the fact that the data loss has a smaller impact here, allowing the baseline to perform better and (ii) the lack of correlated activity in SFHH and HT09 with respect to the LSCH data set, making the random baseline based on the average activity perform better.

[1] N. D. Lane, E. Miluzzo, H. Lu, D. Peebles, T. Choudhury, and A. T. Campbell, A survey of mobile phone sensing, IEEE Commun. Mag. **48**, 140 (2010).

[2] E. Miluzzo, N. D. Lane, K. Fodor, R. Peterson, H. Lu, M. Musolesi, S. B. Eisenman, X. Zheng, and A. T. Campbell, in *Proceedings of the Sixth ACM Conference on Embedded Network Sensor Systems* (ACM, New York, 2008), pp. 337–350.

[3] N. Eagle and A. Sandy Pentland, Reality mining: Sensing complex social systems, Pers. Ubiquit. Comput. **10**, 255 (2006).

[4] C. Cattuto, W. Van den Broeck, A. Barrat, V. Colizza, J.-F. Pinton, and A. Vespignani, Dynamics of person-to-person interactions from distributed RFID sensor networks, PLoS ONE **5**, e11596 (2010).

[5] A. Stopczynski, V. Sekara, P. Sapiezynski, A. Cuttone, M. M. Madsen, J. E. Larsen, and S. Lehmann, Measuring large-scale social networks with high resolution, PLoS ONE **9**, e95978 (2014).

[6] M. Salathé, M. Kazandjieva, J. W. Lee, P. Levis, M. W. Feldman, and J. H. Jones, A high-resolution human contact network for infectious disease transmission, Proc. Natl. Acad. Sci. U.S.A. **107**, 22020 (2010).

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

[8] J. M. Read, K. T. D. Eames, and W. J. Edmunds, Dynamic social networks and the implications for the spread of infectious disease, J. R. Soc. Interface **5**, 1001 (2008).

[9] T. Obadia, R. Silhol, L. Opatowski, L. Temime, J. Legrand, A. C. M. Thiébaut, J.-L. Herrmann, E. Fleury, D. Guillemot, P.-Y. Boëlle *et al.*, Detailed contact data and the dissemination of staphylococcus aureus in hospitals, PLoS Comput. Biol. **11**, e1004170 (2015).

[10] N. Voirin, C. Payet, A. Barrat, C. Cattuto, N. Khanafer, C. Régis, B.-a. Kim, B. Comte, J.-S. Casalegno, B. Lina, and P. Vanhems, Combining high-resolution contact data with virological data to investigate influenza transmission in a tertiary care hospital, Infect. Control Hosp. Epidemiol. **36**, 254 (2015).

[11] W. He, Y. Huang, K. Nahrstedt, and B. Wu, in *Proceedings of the Eighth IEEE International Conference on Pervasive Computing and Communications Workshops* (IEEE, Piscataway, 2010), pp. 141–146.

[12] J. L. Schafer and J. W. Graham, Missing data: Our view of the state of the art, Psychol. Methods **7**, 147 (2002).

[13] T. Smieszek, E. U. Burri, R. Scherzinger, and R. W. Scholz, Collecting close-contact social mixing data with contact diaries: Reporting errors and biases, Epidemiol. Infect. **140**, 744 (2012).

[14] V. Ouzienko and Z. Obradovic, Imputation of missing links and attributes in longitudinal social surveys, Mach. Learn. **95**, 329 (2014).

[15] R. Mastrandrea and A. Barrat, How to estimate epidemic risk from incomplete contact diaries data? PLoS Comput. Biol. **12**, e1005002 (2016).

[16] T. Smieszek, S. Castell, A. Barrat, C. Cattuto, P. J. White, and G. Krause, Contact diaries versus wearable proximity sensors in measuring contact patterns at a conference: Method comparison and participants' attitudes, BMC Infect. Dis. **16**, 341 (2016).

[17] A. C. Ghani, C. A. Donnelly, and G. P. Garnett, Sampling biases and missing data in explorations of sexual partner networks for the spread of sexually transmitted diseases, Stat. Med. **17**, 2079 (1998).

[18] G. Kossinets, Effects of missing data in social networks, Soc. Netw. **28**, 247 (2006).

[19] L. E. C. Rocha, N. Masuda, and P. Holme, Sampling of temporal networks: Methods and biases, Phys. Rev. E **96**, 052302 (2017).

[20] M. Génois, C. L. Vestergaard, C. Cattuto, and A. Barrat, Compensating for population sampling in simulations of epidemic spread on temporal contact networks, Nat. Commun. **6**, 8860 (2015).

[21] C. L. Vestergaard, E. Valdano, M. Génois, C. Poletto, V. Colizza, and A. Barrat, Impact of spatially constrained sampling of temporal contact networks on the evaluation of the epidemic risk, Eur. J. Appl. Math. **27**, 941 (2016).

[22] M. Huisman, Imputation of missing network data: some simple procedures, J. Soc. Struct. **10**, 1 (2009).

[23] L. Lü, L. Pan, T. Zhou, Y.-C. Zhang, and H. E. Stanley, Toward link predictability of complex networks, Proc. Natl. Acad. Sci. U.S.A. **112**, 2325 (2015).

[24] A. Clauset, C. Moore, and M. E. J. Newman, Hierarchical structure and the prediction of missing links in networks, Nature (London) **453**, 98 (2008).

[25] R. Guimerà and M. Sales-Pardo, Missing and spurious interactions and the reconstruction of complex networks, Proc. Natl. Acad. Sci. U.S.A. **106**, 22073 (2009).

[26] A. Godoy-Lorite, R. Guimerà, C. Moore, and M. Sales-Pardo, Accurate and scalable social recommendation using mixed-membership stochastic block models, Proc. Natl. Acad. Sci. U.S.A. **113**, 14207 (2016).

[27] M. Kim and J. Leskovec, in *Proceedings of the 2011 SIAM International Conference on Data Mining* (SIAM, Philadelphia, 2011), Vol. 11, pp. 47–58.

[28] E. Sadikov, M. Medina, J. Leskovec, and H. Garcia-Molina, in *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining* (ACM, New York, 2011), pp. 55–64.

[29] J. Fournet and A. Barrat, Estimating the epidemic risk using non-uniformly sampled contact data, Sci. Rep. **7**, 9975 (2017).

[30] S. Ghonge and D. C. Vural, Inferring network structure from cascades, Phys. Rev. E **96**, 012319 (2017).

[31] M. Nadini, K. Sun, E. Ubaldi, M. Starnini, A. Rizzo, and N. Perra, Epidemic spreading in modular time-varying networks, Sci. Rep. **8**, 2352 (2018).

[32] T. G. Kolda and B. W. Bader, Tensor decompositions and applications, SIAM Rev. **51**, 455 (2009).

[33] L. Gauvin, A. Panisson, and C. Cattuto, Detecting the community structure and activity patterns of temporal networks: A non-negative tensor factorization approach, PLoS ONE **9**, e86028 (2014).

[34] A. Schein, J. Paisley, D. M. Blei, and H. Wallach, in *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (ACM, New York, 2015), pp. 1045–1054.

[35] J. Liu, P. Musialski, P. Wonka, and J. Ye, Tensor completion for estimating missing values in visual data, IEEE Trans. Pattern Anal. Mach. Intell. **35**, 208 (2013).

[36] D. M. Dunlavy, T. G. Kolda, and E. Acar, Temporal link prediction using matrix and tensor factorizations, ACM Trans. Knowl. Discov. Data **5**, 1 (2011).

[37] D. Hric, T. P Peixoto, and S. Fortunato, Network Structure, Metadata, and the Prediction of Missing Nodes and Annotations, Phys. Rev. X **6**, 031038 (2016); R. Bro and S. De Jong, A fast non-negativity-constrained least squares algorithm, J. Chemom. **11**, 393 (1997).

[38] E. Acar, D. M. Dunlavy, T. G. Kolda, and M. Mørup, in *Proceedings of the 2010 SIAM International Conference on Data Mining* (SIAM, Philadelphia, 2010), pp. 701–712.

[39] J.-P. Royer, N. Thirion-Moreau, and P. Comon, in *Proceedings of the 20th European Signal Processing Conference* (Elsevier, Amsterdam, 2012), pp. 1–5.

[40] H. Kim, H. Park, and L. Eldén, in *Proceedings of the Seventh IEEE International Conference on Bioinformatics and Bioengineering* (IEEE, Piscataway, 2007), pp. 1147–1151.

[41] K. Balasubramanian, J. Kim, A. Puretskiy, M. Berry, and H. Park, A fast algorithm for nonnegative tensor factorization using block coordinate descent and an active-set-type method, in *SIAM International Conference on Data Mining, 2010*, Text Mining Workshop (SIAM, Philadelphia, 2010).

[42] E. Acar, M. Nilsson, and M. Saunders, A flexible modeling framework for coupled matrix and tensor factorizations, in *Proceedings of the 22nd European Signal Processing Conference (EUSIPCO), 2014* (IEEE, Piscataway, NJ, 2014), pp. 111–115.

[43] J. Kim and H. Park, in *High-Performance Scientific Computing: Algorithms and Applications*, edited by M. W. Berry, K. A. Gallivan, E. Gallopoulos, A. Grama, B. Philippe, Y. Saad, and F. Saied (Springer, London, 2012), pp. 311–326.

[44] M. Fang, G. Yue, and Q. Yu, The study on an application of Otsu method in Canny operator, *Proceedings of the 2009 International Symposium on Information Processing (ISIP'09) Huangshan, P.R. China, August 21–23, 2009* (Academy Publisher, Finland, 2009), pp. 109–112.

[45] www.sociopatterns.org.

[46] M. Szomszor, C. Cattuto, W. Van den Broeck, A. Barrat, and H. Alani, in *The Semantic Web: Research and Applications*, edited by L. Aroyo, G. Antoniou, E. Hyvönen, A. ten Teije, H. Stuckenschmidt, L. Cabral, and T. Tudorache, Lecture Notes in Computer Science Vol. 6089 (Springer, Berlin, 2010), pp. 196–210.

[47] J. Stehlé, N. Voirin, A. Barrat, C. Cattuto, L. Isella, J.-F. Pinton, M. Quaggiotto, W. Van den Broeck, C. Régis, B. Lina, and P. Vanhems, High-resolution measurements of face-to-face contact patterns in a primary school, PLoS ONE **6**, e23176 (2011).

[48] R. Bro and H. A. L. Kiers, A new efficient method for determining the number of components in PARAFAC models, J. Chemom. **17**, 274 (2003).

[49] A. Vazquez, B. Racz, A. Lukacs, and A.-L. Barabasi, Impact of Non-Poissonian Activity Patterns on Spreading Processes, Phys. Rev. Lett. **98**, 158702 (2007).

[50] J. L. Iribarren and E. Moro, Impact of Human Activity Patterns on the Dynamics of Information Diffusion, Phys. Rev. Lett. **103**, 038702 (2009).

[51] T. Smieszek, A mechanistic model of infection: why duration and intensity of contacts should be included in models of disease spread, Theor. Biol. Med. Modell. **6**, 25 (2009).

[52] J. Stehlé, N. Voirin, A. Barrat, C. Cattuto, V. Colizza, L. Isella, C. Régis, J.-F. Pinton, N. Khanafer, W. Van den Broeck, and P. Vanhems, Simulation of an SEIR infectious disease model on the dynamic contact network of conference attendees, BMC Med. **9**, 87 (2011).

[53] A. Machens, F. Gesualdo, C. Rizzo, A. E. Tozzi, A. Barrat, and C. Cattuto, An infectious disease model on empirical networks of human contact: Bridging the gap between dynamic network data and contact matrices, BMC Infect. Dis. **13**, 185 (2013).

[54] L. Gauvin, A. Panisson, C. Cattuto, and A. Barrat, Activity clocks: Spreading dynamics on temporal networks of human contact, Sci. Rep. **3**, 3099 (2013).

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

[56] A. Barrat and C. Cattuto, in *Social Phenomena*, edited by B. Gonçalves and N. Perra (Springer International, Cham, 2015), Chap. 3, pp. 37–57.