Influence maximization on temporal networks

Şirag Erkol, Dario Mazzilli, and Filippo Radicchi*
Center for Complex Networks and Systems Research, Luddy School of Informatics, Computing, and Engineering, Indiana University, Bloomington, Indiana 47408, USA

(Received 2 July 2020; accepted 14 September 2020; published 19 October 2020)

We consider the optimization problem of seeding a spreading process on a temporal network so that the expected size of the resulting outbreak is maximized. We frame the problem for a spreading process following the rules of the susceptible-infected-recovered model with temporal scale equal to the one characterizing the evolution of the network topology. We perform a systematic analysis based on a corpus of 12 real-world temporal networks and quantify the performance of solutions to the influence maximization problem obtained using different level of information about network topology and dynamics. We find that having perfect knowledge of the network topology but in a static and/or aggregated form is not helpful in solving the influence maximization problem effectively. Knowledge, even if partial, of the early stages of the network dynamics appears instead essential for the identification of quasioptimal sets of influential spreaders.

DOI: 10.1103/PhysRevE.102.042307

I. INTRODUCTION

Influence maximization is a classical optimization problem in network science [1,2]. The problem consists in finding the set of initial spreaders that maximizes the expected outbreak size of a spreading process occurring on a network. The problem is generally solved for a fixed size of the set of initial spreaders. The size of the set is small if compared to the one of the network but large enough to forbid the use of brute-force algorithms for finding exact solutions to the problem. Most of the research on the topic is devoted to the development of approximate algorithms aiming at finding good solutions in a computationally feasible manner. Examples include greedy optimization techniques with relatively high computational complexity but guaranteed performance that can be applied to medium-size networks [2–7] and a multitude of approximate algorithms based on network centrality metrics applicable to large-scale graphs, e.g., Refs. [8–14]. For a systematic analysis of several methods for the identification of influential spreaders in networks see Ref. [15].

Influence maximization has been traditionally studied on static graphs. However, several real-world networks display nontrivial edge temporal variability [16]. If structural variations happen on a timescale comparable with the one of the spreading dynamics, then the two processes interact in a highly nontrivial manner [17–20]. Most of the work in the area of spreading processes on time-varying networks has been focusing on the characterization of their critical properties. Some attention has been devoted to the problem of influence maximization. Specifically, Habiba and Berger-Wolf [21] extend the work by Kempe *et al.* [2] to temporal networks for the susceptible-infected (SI) and linear threshold (LT) models. In their modeling framework, the optimization problem consists in maximizing of a submodular function, thus appearing similar to the one valid for static graphs. However, they show also that ignoring the time order of the interactions generate results that do not relate to the ground-truth spreading process taking place on the temporal network. Similar conclusions are reached by Osawa *et al.*, who study the problem of influence maximization on temporal networks for the SI model [22]. Michalski *et al.* model the temporal network as a multilayer network and the spreading dynamics using the LT model [23]. They analyze solutions to the influence maximization problem under different granularities for the temporal network, including time-aggregated versions of the network. Murata *et al.* propose heuristic methods to solve the influence maximization problem on temporal networks [24]. Han *et al.* [25] and Zhuang *et al.* [26] propose a method where it is assumed that only the topology of the first layer of a temporal network is known. The topology of the succeeding temporal layers can only be discovered based on partial probing of nodes in the network, and such partial topological information is used to select influential spreaders. Gayraud *et al.* focus on the independent cascade (IC) and LT models in a theoretical study about the properties of the influence maximization problem [27]. At odds with many of the influence maximization problems considered in the literature on both static and temporal networks, they show that their optimization problem does not necessarily involve the maximization of a submodular function. They further demonstrate that delaying the activation of some initial spreaders may increase their effective influence.

In this paper, we introduce a discrete-time version of the susceptible-infected-recovered (SIR) model on temporal networks [28]. We systematically study the influence maximization problem associated with SIR spreading on 12 real-world temporal networks. We test the performance of different approximate algorithms aimed at the identification of the best spreaders in the network. Approximations rely

*filiradi@indiana.edu

on different levels of dynamical and topological information. Results of our analysis show that having knowledge, even if partial, of the early stages of the network dynamics is essential for an effective prediction of the influential spreaders in temporal networks.

We stress that our modeling framework is very similar to the one previously used by Valdano *et al.* for SIS spreading on temporal networks [20]. The properties of the two spreading models are, however, rather different, especially because the SIR model displays a sensitivity to the temporal ordering of the network edges that is stronger than the one observed for the SIS model. Further, both the SI and IC models, previously studied by Habiba and Berger-Wolf [21] and Gayraud *et al.* [27], respectively, can be seen as two extreme cases of our model. We extend and generalize those analyses in several respects. First, being able to tune our model between the two extremes, we are essentially able to change the effective level of submodularity of the function that we want to optimize. Second, we do not focus on specific values of the spreading probability for every network. Rather, we tune each network close to its own critical point and study the influence maximization problem near criticality. While studying the phase diagram of the SIR model, we show that the system behavior can be reasonably well predicted by a mean-field approximation and that such an approximation can be effectively used to solve the influence maximization problem. Finally, we do not assume that the optimization problem is solved using full information about the system. Rather, we systematically test the performance of approximations obtained under partial knowledge of the network topology and dynamics [29]. Our results, obtained on a corpus of real-word temporal networks that is significantly larger than those typically considered in previous studies, provide clear indications on the type of ingredients that one needs to rely on when available topological information is incomplete or noisy.

II. EMPIRICAL DATA AND THEIR REPRESENTATION AS TEMPORAL NETWORKS

We focus our attention on 12 empirical datasets containing time-stamped social interactions among pairs of individuals. Datasets refer to two types of interactions. In some of the datasets, interactions correspond to physical proximity contacts, e.g., among high school students [30,31], conference attendees [32], and hospital staff/patients [33]. In other datasets, interactions stand for emails exchanged by coworkers [34]. In all cases, we treat contacts as undirected. All datasets considered in the study are listed in Table I.

Given a dataset, we follow a quite common modeling scheme [16,35,36]. We slice the dataset into time windows of identical length W. We aggregate all interactions within a time slice to form a temporal network layer. Multiple interactions, between a pair of nodes in the same slice, are reduced to a single unweighted link. All layers contain exactly N nodes, where N is the number of distinct individuals involved in at least one social interaction in the dataset. By construction, some nodes may have degree equal to zero in one or more temporal layers. To avoid for the presence of layers that are too sparsely connected, we exclude from our analysis network layers containing a number of null-degree nodes greater than $0.9\,N$.[1] After the cleaning procedure is performed, the dataset is transformed into T temporal network layers. The T layers are chronologically ordered from 1 to T.

TABLE I. Real-world temporal networks. List of the empirical datasets used to construct temporal networks. From left to right, we report the name of the dataset, the length W of the temporal window used to slice the data (time is expressed in seconds), the number T of network layers resulting after slicing and cleaning data, the number of nodes N in the network, and the reference to the paper(s) where the data were first considered.

Dataset	W	T	N	Ref.
Email, dept. 1	2,880,000	18	309	[34]
Email, dept. 2	2,880,000	18	162	[34]
Email, dept. 3	2,880,000	18	89	[34]
Email, dept. 4	2,880,000	18	142	[34]
High school, 2011	14,400	11	126	[38]
High school, 2012	14,400	21	180	[38]
High school, 2013	14,400	14	327	[39]
Hospital ward	14,400	20	75	[33]
Hypertext, 2009	14,400	11	113	[32]
Primary school	7,200	11	242	[30,31]
Workplace	28,800	20	92	[40]
Workplace-2	28,800	20	217	[41]

We choose different W values depending on the dataset on purpose. Our goal is simply ending up with a similar number T of layers across datasets. Also, the threshold value used for disregarding low-density layers is arbitrarily chosen. We are aware that both ingredients affect the construction of the network layers and the outcome of the spreading process taking place on them. We stress, however, that the goal of the paper is not understanding the dynamics of a specific temporal network and/or a specific choice of the parameters of the spreading model. Rather, we want to study the general problem of influence maximization on temporal networks and compare different strategies to solve the problem. As far as we are concerned, the real-world datasets considered here just provide useful data for the construction of temporal network topologies, and influence maximization strategies are compared one against the other on the same test sets.

At the end of the above-described procedure, we have at our disposal a sequence $\{A^{(1)}, \ldots, A^{(t)}, \ldots, A^{(T)}\}$ of T temporal adjacency matrices. The adjacency matrix $A^{(t)}$ fully encodes information about the topology of the tth temporal network layer, with its generic element $A^{(t)}_{ij} = A^{(t)}_{ji} = 1$ if a connection exists between nodes i and j at stage t of the network dynamics, whereas $A^{(t)}_{ij} = A^{(t)}_{ji} = 0$, otherwise.

[1] In some datasets, a large portion of layers, corresponding to periods of inactivity, is excluded from the analysis [37]. For example, in the high school temporal data, there are no recorded night-time interactions, and the layers corresponding to such time frames are excluded as they do not contain edges.

III. SPREADING DYNAMICS

We study spreading dynamics taking place on temporal networks. In analogy with the work by Valdano *et al.* [20], we assume that the characteristic timescales of the spreading process and of the network evolution are identical. We consider the discrete-time version of the SIR model to mimic spreading dynamics [28], thus differentiating from the work by Valdano *et al.* where the SIS model was considered instead. In the SIR model, the state $\sigma_i^{(t)}$ for node i at time t can be $\sigma_i^{(t)} = S, I$, or R. Every node i that is infected at time t, i.e., $\sigma_i^{(t)} = I$, attempts to infect all its susceptible neighbors, i.e., all j such that $A_{ij}^{(t)} = 1$ and $\sigma_j^{(t)} = S$. Infection is successfully transmitted with probability λ. In case of a successful attempt, the state of the newly infected node j changes as $\sigma_j^{(t)} = S \to \sigma_j^{(t+1)} = I$, meaning that the node can spread the infection from time $t+1$ on. After all spreading attempts have been performed, each infected node i may recover with probability μ. A successful recovery attempt changes the state of node i as $\sigma_i^{(t)} = I \to \sigma_i^{(t+1)} = R$. A recovered node does no longer participate in the dynamics, thus it cannot spread nor receive the infection. After all recovery attempts have been performed, time increases as $t \to t+1$.

Two standard models of spreading are obtained as special cases of our more general model. If $\mu = 1$, the model reduces to the IC model [27]. If $\mu = 0$ and no nodes are initially in the recovered state, then the SIR model reduces to the discrete-time version of the SI model [21].

Starting from a given initial configuration $\vec{\sigma}^{(1)} = [\sigma_1^{(1)}, \ldots, \sigma_N^{(1)}]^\intercal$, we follow the dynamics of the model until the last iteration of spreading is performed, reaching the final configuration $\vec{\sigma}^{(T+1)}$. Even for fixed values of λ and μ, the outcome of the spreading model is highly sensitive to the initial conditions (see Fig. 1).

We restrict our attention to special types of initial configurations where all nodes are in the S state, with the exception of a set $\mathcal{X} = \{x_1, x_2, \ldots, x_{|\mathcal{X}|}\}$ of seed nodes that are in the I state, i.e., $\sigma_i^{(1)} = I$ if $i \in \mathcal{X}$ and $\sigma_i^{(1)} = S$ if $i \notin \mathcal{X}$. Given the parameters of the SIR model and the topology of the temporal network, we estimate the relative outbreak size $O(\mathcal{X})$ generated by the seed set $\mathcal{X}$ in a single realization of the SIR model. $O(\mathcal{X})$ is defined as the total number of nodes found either in the R or I state in the stage $T+1$ of the process, divided by the network size N, i.e.,

$$O(\mathcal{X}) = \frac{1}{N} \sum_{i=1}^{N} \left[\mathbb{1}_{\sigma_i^{(T+1)},I} + \mathbb{1}_{\sigma_i^{(T+1)},R} \right], \tag{1}$$

with $\mathbb{1}_{x,y}$ identity operator, i.e., $\mathbb{1}_{x,y} = 1$ if $x = y$ and $\mathbb{1}_{x,y} = 0$ otherwise. $O(\mathcal{X})$ is a random variable, obeying some probability distribution. We stress that $O(\mathcal{X})$ strongly depends on the choice of the parameters λ and μ, the topology of the network layers and their time ordering. However, we do not report the explicit dependence on these factors for shortness of notation.

As in many of the papers on influence maximization [2,15], we use the average value of the outbreak size as the metric of influence for the seed set $\mathcal{X}$. Specifically, we numerically estimate the influence of the seed set $\mathcal{X}$ over a finite number

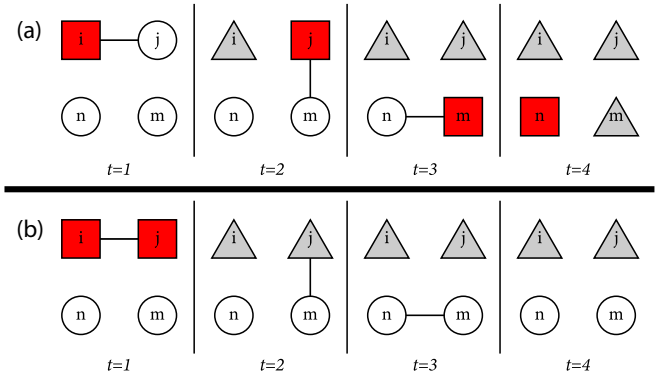

FIG. 1. Susceptible-infected-recovered model on temporal networks. Illustrative example of the modeling framework proposed in this paper, where SIR spreading occurs on a temporal network. In the example, the network consists of four nodes and four temporal layers, and the spreading dynamics takes place over four discrete temporal stages. For simplicity, in the illustration we set the SIR model parameters $\lambda = \mu = 1$ so that the dynamics is deterministic. (a) The initial condition is such that only node i is infected, while all others are in the susceptible state. At the end of the dynamics, all nodes are either infected or recovered. (b) Nodes i and j are initially infected, and they are recovered in the final configuration. Nodes n and m remain in the susceptible state.

Q of numerical simulations as

$$\langle O(\mathcal{X}) \rangle = \frac{1}{Q} \sum_{q=1}^{Q} O_q(\mathcal{X}), \tag{2}$$

where $O_q(\mathcal{X})$ is the relative outbreak size of Eq. (1) obtained in the qth instance of the model. We use $Q = 2,000$ in all our numerical results, unless otherwise specified.

In Fig. 2, we show typical phase diagrams obtained for seed sets of size one. In the diagrams, a data point of the outbreak size for a given pair of parameter values μ and λ is obtained as follows. We consider N different initial conditions, each corresponding to one of the nodes selected as the initial spreader with all other nodes initially in the susceptible state, i.e., $\mathcal{X} = \{i\}$ for all $i = 1, \ldots, N$. For each initial condition, we run $Q = 500$ simulations and estimate the influence of the node according to Eq. (2). We finally take the average value of the influence over all initial conditions as representative quantity for the system outbreak size. The system is characterized by a phase transition from a nonendemic to an endemic phase as the parameters of the spreading model are varied. Specifically, the endemic phase is obtained for sufficiently large values of the spreading probability λ. The critical λ value, namely λ_c, where the transition occurs is a function of the recovery probability μ, i.e., $\lambda_c = \lambda_c(\mu)$. We note that λ_c increases as μ increases. We stress that the system size is finite here, so we are not facing a genuine phase transition. Nonetheless, the change in the value of average outbreak size lets us easily notice the presence of a regime where the outbreak is confined to a small part of the network and a regime where spreading involves a large portion of the system.

We estimate the critical value of the spreading probability $\lambda_c(\mu)$ for a given value of the recovery probability μ as the λ value that maximizes the ratio between standard deviation

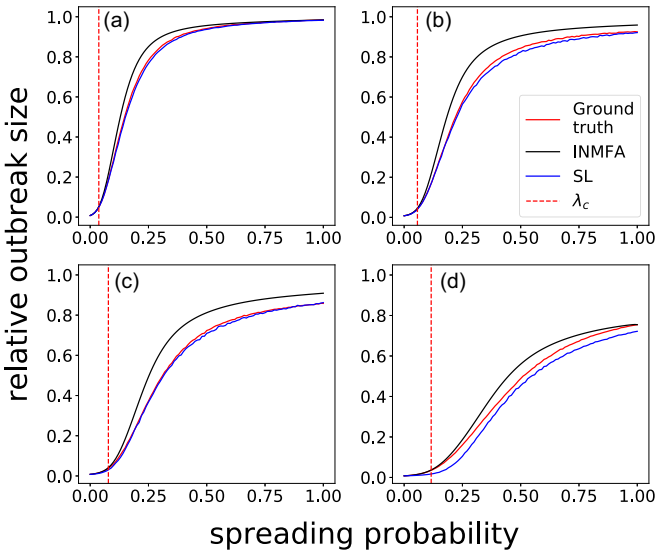

FIG. 2. Epidemic transition in real-world temporal networks. (a) Average value of the relative outbreak size $\langle O(\mathcal{X})\rangle$ as a function of the spreading probability λ. The seed set corresponds to one randomly chosen node. Results are obtained on the "High school, 2011" network and by setting $\mu = 0$. Results from numerical simulations on the real network topology (red curve) are compared against those predicted by INFMA (black curve). The dashed red line indicates the position of our best estimate of the critical value of the spreading probability, i.e., λ_c. We further display results of numerical simulations obtained on the same network topology but with the order of the temporal network layers randomized (SL, blue curve). (b) Same as in (a) but for $\mu = 0.25$. (c) Same as in (a) but for $\mu = 0.5$. (d) Same as in panel (a) but for $\mu = 1$.

and average value of the outbreak size, both computed over $Q = 500$ numerical simulations of the spreading process initiated by a single randomly chosen seed (we use the same procedure as described above, but for simplicity, only nodes with at least one connection in the first layer of the network are considered as possible seeds). We note that looking at the peak of the ratio standard deviation over average value is not the only possible way of defining and identifying the critical point of the transition. One, for example, may look at the position of the peak of the standard deviation only. In general, different definitions may lead to slightly different estimates of λ_c. We stress, however, that the λ_c values we obtain seem to identify quite accurately the transition point (see Fig. 2) and that the exact value of the transition point is not very crucial for the type of analysis we are performing in this paper. In Table II, we report the λ_c values obtained for the various networks considered in this paper.

We note that systems display sensitivity to the temporal organization of the underlying networks, and the sensitivity is more apparent for large μ values than for small μ values. Phase diagrams, and resulting values of the spreading probability where transitions occur, may dramatically vary by simply randomizing the order of the layers but without changing the actual network topology of the layers (see Figs. 2 and 3 and Ref. [37]). This is a quite remarkable difference with respect to the SIS modeling framework by Valdano *et al.*, where the actual order of the temporal layers is not as important for the outcome of the spreading dynamics [20].

TABLE II. Critical thresholds of real-world temporal networks. We report our numerical estimates of the critical spreading probability $\lambda_c(\mu)$ for the temporal networks of Table I. Different columns correspond to different values of the recovery probability μ. Errors associated to the estimates are all equal to or smaller than 10^{-3} and they are not reported in the table for sake of compactness.

Network	$\lambda_c(\mu = 0)$	$\lambda_c(\mu = 0.25)$	$\lambda_c(\mu = 0.5)$	$\lambda_c(\mu = 1)$
Email, dept. 1	0.016	0.043	0.069	0.130
Email, dept. 2	0.010	0.027	0.049	0.096
Email, dept. 3	0.016	0.038	0.066	0.123
Email, dept. 4	0.010	0.029	0.047	0.099
High school, 2011	0.037	0.057	0.078	0.116
High school, 2012	0.025	0.077	0.136	0.205
High school, 2013	0.023	0.042	0.064	0.119
Hospital ward	0.017	0.048	0.087	0.207
Hypertext, 2009	0.023	0.041	0.060	0.097
Primary school	0.013	0.019	0.029	0.043
Workplace	0.042	0.123	0.241	0.308
Workplace-2	0.023	0.063	0.119	0.248

IV. INDIVIDUAL-NODE MEAN-FIELD APPROXIMATION

We can provide a relatively simple description of the spreading process using the individual-node mean-field approximation (INMFA) [28]. The approximation consists in describing the stochastic state variable $\sigma_i^{(t)}$ for node i at time t with the three deterministic variables $S_i^{(t)}$, $I_i^{(t)}$, and $R_i^{(t)}$. Specifically, we have that $S_i^{(t)} = \text{Prob.}[\sigma_i^{(t)} = S]$, i.e., the probability to find node i in state S at stage t over an infinite number of realizations of the process. Similarly, we have that $I_i^{(t)} = \text{Prob.}[\sigma_i^{(t)} = I]$ and $R_i^{(t)} = \text{Prob.}[\sigma_i^{(t)} = R]$. The three deterministic variables are related by the constraint $I_i^{(t)} + S_i^{(t)} + R_i^{(t)} = 1$. Further, the approximation consists in neglecting dynamical correlations among two or more state variables, so that joint probabilities can be replaced by products among marginal probabilities. For example, the probability at stage t of the dynamics to find nodes i and j, respectively, in the infected and recovered states is simply written as $\text{Prob.}[\sigma_i^{(t)} = I, \sigma_j^{(t)} = R] = I_i^{(t)} R_j^{(t)}$.

Under INMFA, we can describe SIR dynamics on the temporal network with the following set of coupled equations:

$$I_i^{(t)} = (1-\mu) I_i^{(t-1)} + \left(1 - I_i^{(t-1)} - R_i^{(t-1)}\right) \times \left[1 - \prod_j \left(1 - \lambda A_{ji}^{(t-1)} I_j^{(t-1)}\right)\right] \tag{3}$$

and

$$R_i^{(t)} = R_i^{(t-1)} + \mu I_i^{(t-1)}. \tag{4}$$

The initial conditions are suitably chosen depending on the problem at hand. In our case, we set $I_i^{(1)} = 1$ for every $i \in \mathcal{X}$ and $S_i^{(1)} = 1$ for $i \notin \mathcal{X}$, and use the above equations to obtain solutions for $t > 1$. Equation (3) simply tells us that the probability that node i is in the infected state at time t is the sum

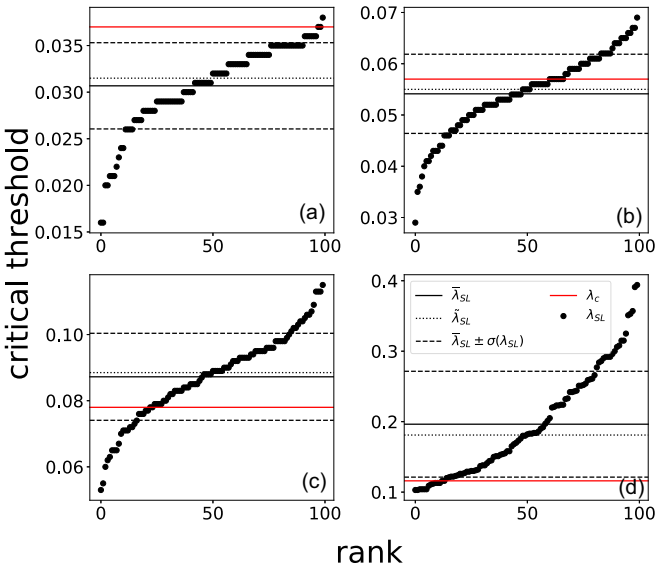

FIG. 3. Sensitivity of the spreading outcome to network dynamics. (a) Best estimates of the critical spreading probability λ_{SL} for randomized versions of the "High school, 2011" temporal network. SIR recovery probability is $\mu = 0$. The randomization consists in reordering the temporal layers only, while the topology of the individual layers is kept invariant. Each black circle corresponds to a specific realization of the randomization process. In the visualization, we simply sort the various realizations depending on their λ_{SL} value. We display horizontal lines identifying the average $\bar{\lambda}_{SL}$ (full black line), the region corresponding to one standard deviation away from the mean [$\bar{\lambda}_{SL} \pm \sigma(\lambda_{SL})$, dashed black lines], the median value $\tilde{\lambda}_{SL}$ (dotted black line), and the actual critical value λ_c measured on the nonrandomized version of the network (red full line, Table II). (b) Same as in (a) but for $\mu = 0.25$. (c) Same as in (a) but for $\mu = 0.5$. (d) Same as in (a) but for $\mu = 1$.

of two terms: (i) The probability that the node was already in the infected state at the previous stage of the dynamics but did not recover and (ii) the probability that the node was not already infected but received the infection by at least one of its infected neighbors at the previous time step. Equation (4) instead tells us that the probability that node i is in the recovered state at time t is the sum of the probability that the node was already recovered or just recovered at the previous stage of the dynamics. INMFA neglects dynamical correlation between nodes. Variables are treated as dynamically independent when instead they are not. In particular, there is a nonnull probability that spreading may occur simultaneously in opposite directions along the same edge, thus causing a systematic overestimation of the true probability of infection. Starting from the imposed initial conditions, one iterates Eqs. (3) and (4) to obtain the marginal probabilities of all nodes in the network at a given stage of the dynamics. The relative size of the outbreak at time t is obtained by simply taking the sum

$$O^{(t)}_{\mathrm{INMFA}} = \frac{1}{N} \sum_{i=1}^{N} \left[I_i^{(t)} + R_i^{(t)} \right]. \tag{5}$$

In Fig. 2, we compare results from the INMFA with ground-truth values obtained from numerical simulations of the spreading process. Due to the independence among variables that is assumed in INMFA, the approximation overestimates the true outbreak size, and under INMFA the phase transition is expected to happen earlier than in the true dynamical system. Nonetheless, we note that INMFA provides a relatively good prediction of the true system outcome, especially in the subcritical regime and around criticality.

V. INFLUENCE MAXIMIZATION

We consider the classical problem of influence maximization consisting in finding the set of seed nodes that maximizes the average size of the outbreak [1,2]. The maximization problem is solved with a constraint on the size of the seed set, and for a given choice of the spreading parameters λ and μ.

A. Greedy optimization

As influence maximization is a NP-hard problem [2], its solutions can only be approximated. On static networks, the best available strategy to solve the problem of influence maximization consists of a greedy algorithm [2,15]. The mechanism of the algorithm is quite simple and naturally extends to temporal networks [27]. Indicated with $\mathcal{B}_v = \{b_1, b_2, \ldots, b_v\}$ the seed set identified by the algorithm at stage v, we initialize the algorithm with $\mathcal{B}_0 = \emptyset$. Then, for $v > 0$ we have

$$b_v = \arg \max_{x \notin \mathcal{B}_{v-1}} \langle O(\mathcal{B}_{v-1} \cup \{x\}) \rangle. \tag{6}$$

Essentially, the best seed set is built sequentially by adding one node at a time. The node b_v selected at stage v is the one providing the largest marginal increment of influence to the existing seed set. We stress that, at each stage v of the algorithm, one needs to numerically estimate $\langle O(\mathcal{B}_{v-1} \cup \{x\}) \rangle$ for all $x \notin \mathcal{B}_{v-1}$ in order to select b_v appropriately, and simulations must be run independently for each potential seed set $\mathcal{B}_{v-1} \cup \{x\}$. We remark that the method requires as inputs full topological and dynamical information about the system, including the actual values of the parameters of the spreading model. In the following, we denote solutions of the influence maximization problem obtained via the standard greedy optimization method with the acronym GR.

For the SIR model in static networks, it is known that influence is a growing and submodular function [2], thus greedy solutions are guaranteed to be within a margin $1 - 1/e \simeq 0.63$ from the true optimum [42]. On temporal networks, the two above conditions are valid only for the special case $\mu = 0$ [21]. However, for $\mu > 0$, influence may decrease as the system size increases, so that the function may also violate the submodularity property (see Ref. [27] and Fig. 1 for a specific example). As a consequence, the greedy algorithm does not guarantee a known optimality gap.

B. Approximate greedy optimization

The complexity of the algorithm described in Eq. (6) grows cubically with the network size, as estimates of the influence of all seed sets are obtained via numerical simulations. Complexity reduction is possible by approximating $\langle O(X) \rangle$ in some way, so that the elementary choice of Eq. (6) is

replaced by

$$b_v = \arg \max_{x \notin \mathcal{B}_{v-1}} F(\mathcal{B}_{v-1} \cup \{x\}). \quad (7)$$

Here we indicated with $F(\mathcal{B} \cup \{x\})$ a generic function that estimates the incremental importance of node $x \notin \mathcal{B}$ for the influence of the set $\mathcal{B}$, assuming that influence is not directly measured. Typical choices of F leverage parallel and/or partial computation to decrease computational complexity. For the IC model on static networks for example, the equivalence of the spreading model with static bond percolation suggests how to decrease algorithmic complexity without sacrificing performance [4,7]. In Ref. [4], $F(\mathcal{B} \cup \{x\})$ is defined as the average size of the clusters that contain node x but no nodes already in $\mathcal{B}$, a quantity that is equivalent to the targeted ground truth $\langle O(\mathcal{B} \cup \{x\})\rangle$ and that can be computed in parallel for all nodes. Variations of the methods of Refs. [4,7] are not easily implementable for the general SIR model, and the temporal nature of the network creates additional challenges. We implement, however, an approximate version of greedy optimization that uses the INMFA prediction of Eq. (5) for the definition of the function F. Many other methods aiming at reducing algorithmic complexity use network centrality metrics for the definition of F such that, during the course of algorithm, the score is static (e.g., degree centrality) or can be quickly recomputed with partial computation (e.g., adaptive degree centrality). On the basis of previous analyses conducted on static networks [15], we focus our attention on adaptive degree centrality only.

C. Greedy optimization under incomplete information

In the sections above, we made the strong hypothesis that optimization is performed by knowing in advance that the network is evolving, and how it exactly evolves. Further, the optimization is performed by being aware of the true spreading dynamics, including the actual values of the model parameters and the existence of a specific temporal horizon in the spreading process.

Having full knowledge of all the ingredients of the problem is, however, a strong assumption. In realistic scenarios, it is much more likely to attempt to solve the problem with limited and/or noisy information. For example, we may have at our disposal only a flat and aggregated version of the true network, where temporal information is absent. In this scenario, we would apply the greedy algorithm to a static network, disregarding completely the existence of network evolution and the time horizon for the spreading. We would further be able to identify the critical regime of spreading for the static network only. In essence, we would use still the same approach as Eq. (7) where the function F represents an approximation of the ground-truth $\langle O\rangle$ of Eq. (6). However, the approximation would not be made with the goal of reducing computational complexity. Rather, it would be enforced by the incompleteness of the information at our disposal in the solution of the true problem.

There are many potential ways in which topological information may arrive to us incomplete or noisy. We consider several possibilities listed in Table III. The simplest setting is what we call SL, where layers are randomly reordered, but all other information required for the solution of the influence maximization problem is preserved. SL is the same setting already considered in the study of the sensitivity of the outbreak size to the temporal ordering of the network layers (Figs. 2 and 3). We remark that, in the SL setting, we still rely on the same exact scheme as described for greedy optimization. Thus, to perform the selection step of Eq. (7), we run multiple numerical simulations of SIR dynamics by assuming that we know the true values of the spreading and recovery probabilities of the spreading model, but also that we believe that the true network dynamics is given by the specific SL setting at our hand.

TABLE III. Identification of influential spreaders in temporal networks. We list here the various approximations used in the solution of the influence maximization problem. From left to right, the columns of the table report: the acronym of the approximation, awareness by the approximation about the existence of a temporal horizon in the spreading, awareness by the approximation about the temporal evolution of the network topology, number/type of temporal layers used in the approximation. The various approximations are described in the main text.

Approximation	Time horizon	Time order	Temporal layers
GR	Yes	Yes	All
INMFA	Yes	Yes	All
RND	No	No	None
SL	Yes	No	All
FL	No	No	One
RL	No	No	One
ST	No	No	Aggregate
AD-F	No	No	One
AD-A	No	No	Aggregate

We then consider cases where part of the temporal information is not present in our input data. For example, we consider the setting FL where only the first temporal layer is used in the solution of the problem. The setting RL is analogous to FL with the only difference that one randomly chosen layer is selected to play the role of the static network. In these cases, we perform standard greedy optimization under the hypothesis of having a static network topology, and we assume that the critical value of the spreading probability is given by the one of the static network.

We then consider a scenario where temporal information is flattened. The solution to the influence maximization problem relies on an aggregated version of the network and no temporal horizon is provided in the estimate of the outbreak size. We name this setting ST. We perform standard greedy optimization under the hypothesis of having a static network topology, and we assume that the critical value of the spreading probability is the one valid for the aggregated static network.

Also, we consider further approximations where the optimization problem is solved using adaptive degree (AD) centrality computed using full or partial information of the system topology. AD is a metric similar to degree centrality. The only difference is that, once a node is selected as a seed, the node is considered as removed from the network and the degree values of all other nodes are recomputed before selecting the next seed. Essentially, the function $F(\mathcal{B}_{v-1} \cup \{x\})$

appearing in Eq. (7) equals the number of connections of node x with nodes that do not belong to the set $\mathcal{B}_{v-1}$. Specifically, we consider the approximations AD-F and AD-A where, respectively, the first layer or the aggregation of all layers are used for the computation of the adaptive degree centrality. The method in this case relies on topological information only, and there is no need of feeding the algorithm with information about the spreading model and its parameter values.

Finally, we consider the setting RND, where the seed set is built by randomly selecting nodes. This is the only viable option in case nothing is known about the network, and should provide the worst performance possible.

D. Systematic tests of performance

We apply the different approximations of Table III in the identification of the influential spreaders in the 12 temporal networks constructed from the datasets of Table I. We consider four distinct values of the recovery probability $\mu = 0$, $\mu = 0.25$, $\mu = 0.5$, and $\mu = 1$. For each network and μ value, we consider the critical value $\lambda_c(\mu)$ of the spreading probability, see Table II. We finally consider separately the cases $\lambda = 0.5\,\lambda_c(\mu)$, $\lambda = \lambda_c(\mu)$, and $\lambda = 2\,\lambda_c(\mu)$ as representative for the subcritical, critical, and supercritical dynamical regimes, respectively. In summary, for each of the 12 temporal networks we consider 12 distinct combinations of μ and λ values, so that each approximation for the solution of the influence maximization problem is tested in 144 different experimental settings.

We stress that the true values of the spreading probability are used only for predictions under the GR, INMFA, and SL settings. The other methods assume different critical values for λ, and predictions for the various regimes are made using such a value as a reference. Predictions of all approximations are tested on the ground-truth dynamics. That is, given a set of predicted seeds, we run numerical simulations of the spreading process using true parameter values on the true temporal network.

A typical outcome of the systematic analysis we perform is displayed in Fig. 4. There, we plot the average value of the relative outbreak size as a function of the relative size of the seed set. We display results only for a selection of the approximations listed in Table III, and only for the critical regime of spreading. Results for the other methods, and for other dynamical regimes, are reported in Ref. [37]. The best solution is generally obtained by GR, i.e., the straight implementation of the greedy optimization strategy of Eq. (6). This is not surprising as the method relies on complete topological and dynamical information. Although $\langle O \rangle$ is not submodular, the shape of the curve indicates an effective submodular behavior resulting from the maximization process. Results obtained under INMFA are generally very close to (sometimes even slightly better than) those of GR. As one may expect, the worst performance is obtained by random selection, i.e., RND. Using FL provides a quite good performance despite other temporal information being neglected. SL, where one is aware of network evolution, but does not know exactly how the temporal layers are ordered, displays poor performance.

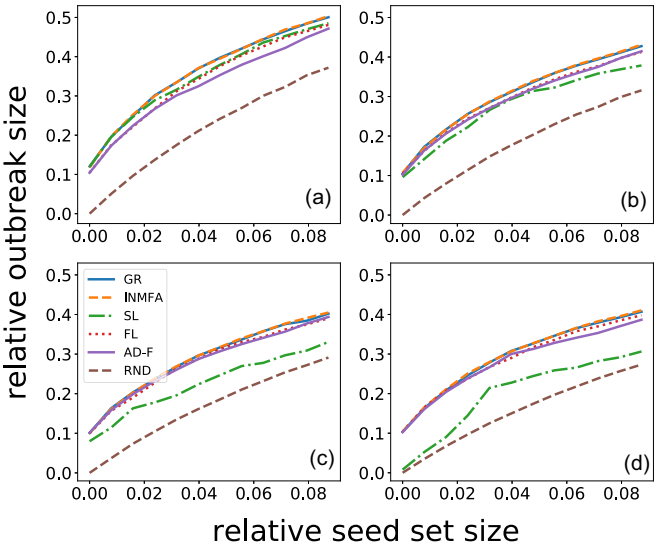

FIG. 4. Identification of influential spreaders in temporal networks. (a) Average value of the relative size of the outbreak, i.e., $\langle O(X) \rangle$ [see Eq. (2)], as a function of the relative size of the seed set, i.e., $|X|/N$. The seed set is selected according to some of the approximations described in the text and listed in Table III. The network analyzed is "High school, 2011." Spreading dynamics is critical, with recovery probability $\mu = 0$ and $\lambda = \lambda_c(\mu) = 0.037$. (b) Same as in (a), but for $\mu = 0.25$ and $\lambda = \lambda_c(\mu) = 0.057$. (c) Same as in (a), but for $\mu = 0.5$ and $\lambda = \lambda_c(\mu) = 0.078$. (d) Same as in (a), but for $\mu = 1$ and $\lambda = \lambda_c(\mu) = 0.116$.

AD-F provides a fair approximation in terms of performance even though it has no information on spreading dynamics.

We use GR as the baseline for assessing the quality of the solutions obtained under the other approximations [15,29]. Given a network with N nodes, we fix the targeted seed set size to $V = 0.1\,N$. For the generic approximation a, we evaluate the area under the curve of the spreading impact of the seed set $\mathcal{B}^{(a)}$ that the approximation identifies, i.e.,

$$\mathrm{AUC}_a = \sum_{v=1}^{V} \langle O(\mathcal{B}_v^{(a)}) \rangle, \tag{8}$$

where $\mathcal{B}_v^{(a)}$ is the seed set found by the approximation a in the vth step of the optimization algorithm of Eq. (7). The definition can be clearly adapted to compute $\mathrm{AUC}_{\mathrm{GR}}$, thus leading to a metric of performance for the straight implementation of the greedy algorithm of Eq. (6). We note that the metric AUC_a gives importance to the global impact of the seed set $\mathcal{B}_V^{(a)}$ but also to the order in which nodes are placed in the set during the optimization steps of Eq. (7). We then normalize the performance p_a of the generic approximation a with GR by simply taking the ratio

$$p_a = \frac{\mathrm{AUC}_a}{\mathrm{AUC}_{\mathrm{GR}}}. \tag{9}$$

In Fig. 5, we report summary results of our systematic analysis. Performance of the various approximations highly depend on both the parameters μ and λ. Many of the approximations reach nearly optimal performances for small μ and λ values. As μ grows, having perfect knowledge of the

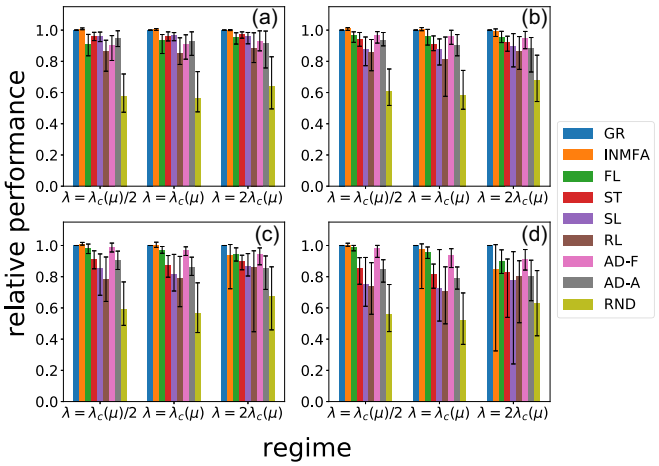

FIG. 5. Identification of influential spreaders in temporal networks. (a) Performance, as defined in Eq. (9), of the various approximations listed in Table III. Performance values are relative to those obtained for GR. The height of the colored bars indicate average values of the relative performance over the set of the twelve temporal networks studied in this paper, see Table I. Error bars identify minimum and maximum values of the performance measured over the entire corpus of real networks. We study different dynamical regimes by selecting different spreading probability values while keeping the recovery probability fixed at $\mu = 0$. (b) Same as in (a) but for $\mu = 0.25$. (c) Same as in (a) but for $\mu = 0.5$. (d) Same as in (a) but for $\mu = 1$.

initial topology of the network, such as in the FL or AD-F approximations, becomes essential to reach good performance. Approximations that do not rely on such a knowledge lose 10–20% in performance compared to the performance displayed by the same approximations at low μ values.

VI. CONCLUSIONS

Irreversible spreading models, such as the SIR model, display outcomes that strongly depend on the initial conditions. Such a sensitivity is apparent in static networks already but gets amplified when the network exhibits temporal changes on a timescale comparable with the one of the spreading dynamics. While seeking solutions to the problem of influence maximization, sensitivity of the spreading outcome to initial conditions is further extremized. Indeed, our systematic analysis shows that good solutions to the influence maximization problem require accurate knowledge of the network dynamics, especially regarding the order in which network edges appear in the system. Also, one of our most important numerical findings is that having knowledge of only the first snapshot of a temporal network is still sufficient for identifying influential spreaders effectively. The topological characteristics of the nodes selected as spreaders depend on the dynamical regime. If the recovery probability μ is large, then nodes that are central in the first few layers are good spreaders. If μ is small instead, then nodes that are central on average over all layers are good spreaders. For example, we see that AD-F outperforms AD-A in all settings except for the subcritical and critical regimes when $\mu = 0$.

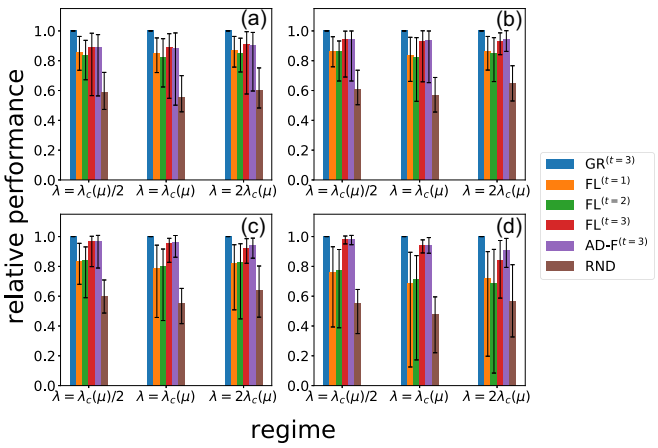

FIG. 6. Prediction of influential spreaders in temporal networks. (a) Same as in Fig. 5(a) with the difference that the ground-truth dynamics is started at time $t = 3$ instead of time $t = 1$. Predictions using the $GR^{(t)}$, $FL^{(t)}$, and $AD\text{-}F^{(t)}$ approximations are based on perfect knowledge of the network topology/dynamics, but under the assumption that spreading starts at time t. (b) Same as in (a) but for $\mu = 0.25$. (c) Same as in (a) but for $\mu = 0.5$. (d) Same as in (a) but for $\mu = 1$.

Our entire analysis is based on the evaluation of the performance of different ex-post approximations. In practical settings, however, it may be more realistic to expect the observer to be aware of past snapshots of a temporal network, and use this information to make predictions about top influencers for a future spreading process taking place on temporal network layers with unknown topology. As an illustrative example, in Fig. 6, we show performance results obtained by comparing *ex ante* predictions of the influential spreaders under approximations similar to those we considered earlier in the paper. Specifically, we name as $FL^{(t)}$ the approximation relying on layer t as the only information available about the network, and with $AD\text{-}F^{(t)}$ the approximation relying on adaptive degree centrality computed on the tth layer of the network. We use layers $t = 1, 2$, and 3 to make predictions about the spreaders in the temporal network. The true dynamics starts from layer $t = 3$ in the simulations. We measure the performance of our predictions relative to the best achievable one, here named as $GR^{(t=3)}$. The results of Fig. 6 show that the lack of information about the initial layer of the dynamics leads to a significant drop in performance. Even a small delay between the last known layer and the start of the process may affect significantly the results. This fact indicates that further research is needed to design effective methods for the prediction of influential spreaders in temporal networks.

ACKNOWLEDGMENTS

S.E. and F.R. acknowledge support from the National Science Foundation (CMMI-1552487). D.M. and F.R. acknowledge support from the US Army Research Office (W911NF-16-1-0104).

[1] P. Domingos and M. Richardson, Mining the network value of customers, in *Proceedings of the 7th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (ACM, New York, 2001), pp. 57–66.

[2] D. Kempe, J. Kleinberg, and É. Tardos, Maximizing the spread of influence through a social network, in *Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (ACM, New York, 2003), pp. 137–146.

[3] J. Leskovec, A. Krause, C. Guestrin, C. Faloutsos, J. VanBriesen, and N. Glance, Cost-effective outbreak detection in networks, in *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (ACM, New York, 2007), pp. 420–429.

[4] W. Chen, Y. Wang, and S. Yang, Efficient influence maximization in social networks, in *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (ACM, New York, 2009), pp. 199–208.

[5] A. Goyal, W. Lu, and L. V. Lakshmanan, Celf + + optimizing the greedy algorithm for influence maximization in social networks, in *Proceedings of the 20th International Conference Companion on World Wide Web* (ACM, New York, 2011), pp. 47–48.

[6] H. T. Nguyen, M. T. Thai, and T. N. Dinh, Stop-and-stare: Optimal sampling algorithms for viral marketing in billion-scale networks, in *Proceedings of the 2016 International Conference on Management of Data* (ACM, New York, 2016), pp. 695–710.

[7] Y. Hu, S. Ji, Y. Jin, L. Feng, H. E. Stanley, and S. Havlin, Local structure can identify and quantify influential global spreaders in large scale social networks, Proc. Natl. Acad. Sci. USA **115**, 7468 (2018).

[8] E. Estrada and J. A. Rodriguez-Velazquez, Subgraph centrality in complex networks, Phys. Rev. E **71**, 056103 (2005).

[9] M. Kitsak, L. K. Gallos, S. Havlin, F. Liljeros, L. Muchnik, H. E. Stanley, and H. A. Makse, Identification of influential spreaders in complex networks, Nat. Phys. **6**, 888 (2010).

[10] G. F. de Arruda, A. L. Barbieri, P. M. Rodríguez, F. A. Rodrigues, Y. Moreno, and L. F. Costa, Role of centrality for the identification of influential spreaders in complex networks, Phys. Rev. E **90**, 032812 (2014).

[11] F. Morone and H. A. Makse, Influence maximization in complex networks through optimal percolation, Nature, **524**, 65 (2015).

[12] L. Lü, D. Chen, X.-L. Ren, Q.-M. Zhang, Y.-C. Zhang, and T. Zhou, Vital nodes identification in complex networks, Phys. Rep. **650**, 1 (2016).

[13] P. Clusella, P. Grassberger, F. J. Pérez-Reche, and A. Politi, Immunization and Targeted Destruction of Networks Using Explosive Percolation, Phys. Rev. Lett. **117**, 208301 (2016).

[14] L. Zdeborová, P. Zhang, and H.-J. Zhou, Fast and simple decycling and dismantling of networks, Sci. Rep. **6**, 37954 (2016).

[15] Ş. Erkol, C. Castellano, and F. Radicchi, Systematic comparison between methods for the detection of influential spreaders in complex networks, Sci. Rep. **9**, 15095 (2019).

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

[17] B. A. Prakash, H. Tong, N. Valler, M. Faloutsos, and C. Faloutsos, Virus propagation on time-varying networks: Theory and immunization algorithms, in *Proceedings of the Joint European Conference on Machine Learning and Knowledge Discovery in Databases* (Springer, Berlin, 2010), pp. 695–710.

[18] M. Karsai, M. Kivelä, R. K. Pan, K. Kaski, J. Kertész, A.-L. Barabási, and J. Saramäki, Small but slow world: How network topology and burstiness slow down spreading, Phys. Rev. E **83**, 025102(R) (2011).

[19] N. Perra, B. Gonçalves, R. Pastor-Satorras, and A. Vespignani, Activity driven modeling of time varying networks, Sci. Rep. **2**, 469 (2012).

[20] E. Valdano, L. Ferreri, C. Poletto, and V. Colizza, Analytical Computation of the Epidemic Threshold on Temporal Networks, Phys. Rev. X **5**, 021005 (2015).

[21] T. Habiba and T. Berger-Wolf, Maximizing the extent of spread in a dynamic network, Technical Report 20 (2007).

[22] S. Osawa and T. Murata, Selecting seed nodes for influence maximization in dynamic networks, in *Complex Networks VI* (Springer, Berlin, 2015), pp. 91–98.

[23] R. Michalski, T. Kajdanowicz, P. Bródka, and P. Kazienko, Seed selection for spread of influence in social networks: Temporal vs. static approach, New Gener. Comput. **32**, 213 (2014).

[24] T. Murata and H. Koga, Extended methods for influence maximization in dynamic networks, Comput. Soc. Netw. **5**, 8 (2018).

[25] M. Han, M. Yan, Z. Cai, Y. Li, X. Cai, and J. Yu, Influence maximization by probing partial communities in dynamic online social networks, Trans. Emerg. Telecommun. Technol. **28**, e3054 (2017).

[26] H. Zhuang, Y. Sun, J. Tang, J. Zhang, and X. Sun, Influence maximization in dynamic social networks, in *Proceedings of the 2013 IEEE 13th International Conference on Data Mining* (IEEE, Los Alamitos, CA, 2013), pp. 1313–1318.

[27] N. T. Gayraud, E. Pitoura, and P. Tsaparas, Diffusion maximization in evolving social networks, in *Proceedings of the 2015 ACM on Conference on Online Social Networks* (ACM, New York, 2015), pp. 125–135.

[28] R. Pastor-Satorras, C. Castellano, P. Van Mieghem, and A. Vespignani, Epidemic processes in complex networks, Rev. Mod. Phys. **87**, 925 (2015).

[29] Ş. Erkol, A. Faqeeh, and F. Radicchi, Influence maximization in noisy networks, Europhys. Lett. **123**, 58007 (2018).

[30] V. Gemmetto, A. Barrat, and C. Cattuto, Mitigation of infectious disease at school: Targeted class closure vs school closure, BMC Infect. Dis. **14**, 695 (2014).

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

[32] L. Isella, J. Stehlé, A. Barrat, C. Cattuto, J.-F. Pinton, and W. Van den Broeck, What's in a crowd? Analysis of face-to-face behavioral networks, J. Theor. Biol. **271**, 166 (2011).

[33] P. Vanhems, A. Barrat, C. Cattuto, J.-F. Pinton, N. Khanafer, C. Régis, B.-a. Kim, B. Comte, and N. Voirin, Estimating potential infection transmission routes in hospital wards using wearable proximity sensors, PLoS ONE **8**, e73970 (2013).

[34] A. Paranjape, A. R. Benson, and J. Leskovec, Motifs in temporal networks, in *Proceedings of the 10th ACM International Conference on Web Search and Data Mining* (ACM, New York, 2017), pp. 601–610.

[35] P. J. Mucha, T. Richardson, K. Macon, M. A. Porter, and J.-P. Onnela, Community structure in time-dependent, multiscale, and multiplex networks, Science **328**, 876 (2010).

[36] M. Kivelä, A. Arenas, M. Barthelemy, J. P. Gleeson, Y. Moreno, and M. A. Porter, Multilayer networks, J. Complex Netw. **2**, 203 (2014).

[37] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevE.102.042307 for a description of data preprocessing and additional analyses.

[38] J. Fournet and A. Barrat, Contact patterns among high school students, PLoS ONE **9**, e107878 (2014).

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

[40] M. Génois, C. L. Vestergaard, J. Fournet, A. Panisson, I. Bonmarin, and A. Barrat, Data on face-to-face contacts in an office building suggest a low-cost vaccination strategy based on community linkers, Netw. Sci. **3**, 326 (2015).

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

[42] G. L. Nemhauser, L. A. Wolsey, and M. L. Fisher, An analysis of approximations for maximizing submodular set functions–i, Math. Program. **14**, 265 (1978).