Sampling of temporal networks: Methods and biases

Luis E. C. Rocha[*]
*Department of Public Health Sciences, Karolinska Institutet, 17177 Stockholm, Sweden
and Department of Mathematics, Université de Namur, 5000 Namur, Belgium*

Naoki Masuda
Department of Engineering Mathematics, University of Bristol, Bristol BS8 1UB, United Kingdom

Petter Holme
Institute of Innovative Research, Tokyo Institute of Technology, Tokyo 152-8550, Japan

(Received 8 August 2017; published 1 November 2017)

Temporal networks have been increasingly used to model a diversity of systems that evolve in time; for example, human contact structures over which dynamic processes such as epidemics take place. A fundamental aspect of real-life networks is that they are sampled within temporal and spatial frames. Furthermore, one might wish to subsample networks to reduce their size for better visualization or to perform computationally intensive simulations. The sampling method may affect the network structure and thus caution is necessary to generalize results based on samples. In this paper, we study four sampling strategies applied to a variety of real-life temporal networks. We quantify the biases generated by each sampling strategy on a number of relevant statistics such as link activity, temporal paths and epidemic spread. We find that some biases are common in a variety of networks and statistics, but one strategy, uniform sampling of nodes, shows improved performance in most scenarios. Given the particularities of temporal network data and the variety of network structures, we recommend that the choice of sampling methods be problem oriented to minimize the potential biases for the specific research questions on hand. Our results help researchers to better design network data collection protocols and to understand the limitations of sampled temporal network data.

DOI: 10.1103/PhysRevE.96.052302

I. INTRODUCTION

Networks have been used to model the interactions and interdependencies between the parts of a system [1]. Social and sexual contacts, flights between airports, email and phone communication, or gene regulatory networks are just a few examples of systems that can be conveniently mapped into networks [1,2]. When modeling real systems as networks, researchers sample data by extracting the relevant information within a given temporal and spatial frame [3], trace-routing or snowballing from one or multiple sources [4,5], or simply by collecting all network-related information of a specific system, for example, email exchanges within a university or social interactions on a web community [2,6]. Sampling network data involves at least four main decisions: the choice of (i) the total observation, or sampling, time (e.g., 1 day or 1 year); (ii) which nodes and (iii) links will be observed (e.g., all or a fraction); and (iv) the temporal resolution, i.e., the time interval in which data are recorded. If the temporal resolution is smaller than the total observation time, then several interaction events between the same pair of nodes may be recorded and filtering strategies may be used to remove weak links [7].

Network modeling may involve the traditional framework of static networks or extensions such as temporal networks [8]. In temporal networks, nodes and links are active at given times in contrast to static networks where nodes and links remain active during the whole period. Temporal networks thus describe more realistically the temporal paths through which information (e.g., through email communication [9]), infections (e.g., over sexual contacts [10]), or resources (e.g., via flights [11]) can propagate or flow. In this temporal perspective, the order and frequency of node and link activations directly affect the dynamics of simulated epidemics [12–17] and information spread [18–21], mixing properties of random walks [16,22–24], and synchronization [25–27] on networks. Although some level of recording error is acceptable, accurate labeling of the interaction events is important to study, for example, simulated infections on real-life temporal networks [28,29].

Another challenge that comes with the study of real-life temporal networks is the amount of generated data since all timings of link activation are stored. This is in contrast to static networks in which activation events are aggregated and multiple activations of the same link are then represented as single weights [30], saving computer memory. The memory cost is particularly problematic when handling big data or when designing studies to collect social interactions using electronic devices such as radio-frequency identification tags [31,32] and mobile phones [33]. In both cases, researchers aim to collect as much relevant data as possible while optimizing resources. Furthermore, several algorithms used to extract information or to simulate dynamic processes on networks struggle to deal with large temporal networks, becoming computationally intractable [34–38]. Facing these challenges, the natural question that emerges is what data should be collected and used in network studies.

The four sampling decisions mentioned above are more critical for the study of temporal networks than for static networks. These decisions are, however, mostly case dependent, meaning that in particular contexts, one or another

[*] luis.rocha@ki.se

factor may weight more, or less, when designing the sampling protocol. For example, the total observation time might affect birth and death statistics of nodes and links and add artificial cutoffs to interevent times since interaction events might be truncated. Similarly, the temporal resolution acts like a filter since only temporal patterns of node and link activity at time scales above the resolution are observable. A typical example is to use a resolution of 1 day to collect data on email communication; this choice would miss the rich dynamic communication patterns happening within a day. The sampling of nodes and links are expected to have at least the same effect as on static networks [5,6] with the aggravated consequence that missing nodes and links would also affect the temporal patterns of the neighboring nodes.

When sampling temporal network data, one wishes to collect as much information as possible such that both short- and long-term temporal patterns can be observed [39]. Yet the amount of information should be manageable by existing algorithms. In this paper, we study the impact of four sampling design decisions, or strategies, on key temporal network variables applied to various categories of real-life temporal networks. In particular, we will study how the choice of the observation time, the temporal resolution, and the number of sampled nodes and events affect the statistics of the lifetime and burstiness of links, the number and length of temporal paths between nodes, and the number of secondary infections and outbreak size of simulated epidemics. The choice of statistics and data sets is not exhaustive and indeed several other options are possible. We focus on statistics that are typically used to characterize temporal activity, paths, and spreading processes and on data sets that are relevant to study human dynamics, particularly epidemic and information spread, in different contexts.

II. MATERIALS AND METHODS

A. Temporal networks

For a given time period T, a temporal network of size N is defined as a set of nodes i connected by a set E of links (i,j), in which events occur at times t [8]. M represents the sum of the number of events in each link for all links. The temporal resolution δ characterizes the size of the time interval (or snapshot) in which the network data are collected; therefore, an event occurring at time t actually means that the event occurred in the time interval $[t,t+\delta)$. The statistics of the sampled networks will be represented by the subscript s, for example, N_s and M_s represent the number of nodes and of events, respectively.

B. Network data

We will use six network data sets corresponding to different contexts in which temporal networks are relevant. We have chosen networks with different topological and temporal structures. The first data set corresponds to sexual contacts between sex-workers and their clients (SEX) [14,40]; the second is about online communication between users of a web-community related to movies (FOR) [41]; the third is about email communication within a university (EMA) [9]; the fourth is about online communication between students in an online social network (COL) [42]; the fifth is about face-to-face proximity contacts (⩽1.5 m) between high-school students (HSC) [43]; the sixth is also about proximity contacts but between visitors of a museum exposition (GAL) [44] (see Table I). Links are undirected and only a single event may occur in a time window $[t,t+\delta)$ for a given link, i.e., events are unweighted.

TABLE I. Summary statistics of the original temporal networks. Number of nodes (N), number of events (M), temporal resolution (δ), and observation time (T).

	N	M	δ	T
SEX	11 416	33 508	1 d	1000 d
FOR	7084	625 435	1 d	3142 d
EMA	3186	234 412	1 h	1959 h
COL	1899	37 178	1 h	4649 h
HSC	310	47 338	20 s	32 360 s
GAL	204	6709	20 s	29 000 s

C. Sampling methods

Sampling consists in making a number of observations or selecting a set of individuals to estimate properties of the target population. In the context of networks, sampling means selecting a number of nodes and links of a system within temporal and spatial frames to build the network of interest. In this paper, we will take network data sets available in the literature as reference populations. We will then study the consequences, on the network structures and on dynamics on the networks, of applying different sampling strategies on these populations. Effectively, we will subsample the original empirical network and then discuss the biased estimates of each sample, that is, the difference in the estimates given by the sampled and the original networks. This subsampling approach is widely used in statistics (see, e.g., subsampling bootstrap [45]) and other disciplines (see, e.g., Refs. [5,6]).

We will study the effect of four sampling strategies (Fig. 1): (i) to reduce the observation time T_s, where $T_s \leqslant T$ and $[0,T_s]$ is the sampling time in which the network data are collected (strategy TS); (ii) to uniformly select a fraction N_s/N of nodes of the original network and thus all events between the sampled nodes (strategy NS); (iii) to uniformly select a fraction M_s/M of events of the original network and thus all nodes connected by these events (strategy ES)—note that this protocol is used, instead of selecting links (and consequently all events associated to that particular link), because of higher flexibility and because one can design “on-line sampling,” that is, collect events as they happen in time; and (iv) to reduce the resolution by setting δ_s a multiple of δ of the original network (strategy RS). Note that repeated same-link events in the interval $[t,t+\delta_s)$ are merged into a single event.

D. Validation measures

To compare the effects of the four sampling strategies, we will estimate six measures, or statistics, on each sample s of the original networks. For strategies NS and ES, we will present average values calculated over five random network samples. Two measures are related to the timings of events, two to the temporal paths and two to the dynamics on the network. These

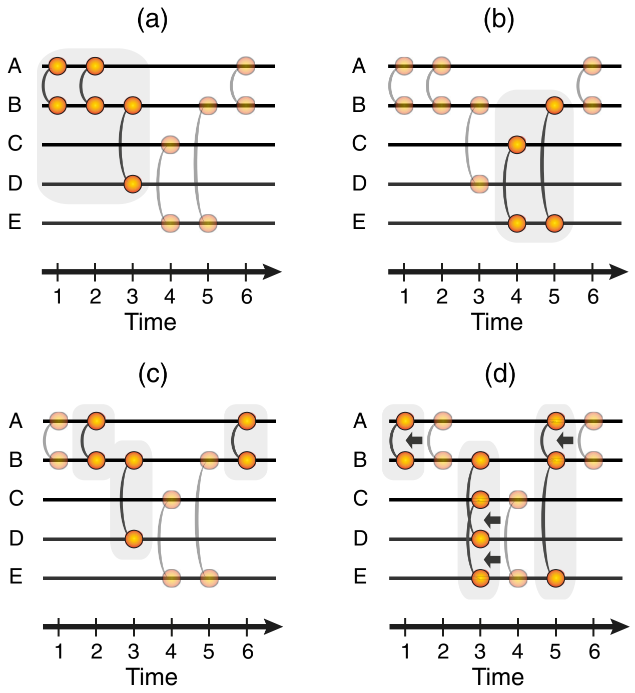

FIG. 1. All panels show a timeline representation of a temporal network where one horizontal line represents a node and there is a vertical line connecting two nodes if they interact at a particular time (i.e., an event). Sampled nodes and sampled events are highlighted for each strategy. In (a), we obtain a new temporal network by truncating the observation time to T_s. In this example $T_s = 3$, and therefore all nodes and events in $1 \leqslant t \leqslant 3$ are collected. In (b), we uniformly choose nodes in $1 \leqslant t \leqslant T$. In this example, nodes B, C, and E are sampled, and therefore only the events between these particular nodes are collected. In (c), we uniformly choose events in $1 \leqslant t \leqslant T$. In this example, the events (A,B) at $t = 2$, (B,D) at $t = 3$, and (A,B) at $t = 6$ are sampled. In (d), we coarse-grain the temporal network in the interval $1 \leqslant t \leqslant T$ by letting an event represent the presence of at least one event at that link during δ_s. In this example, we change the resolution from $\delta = 1$ to $\delta_s = 2$, and therefore we only record interaction events at times 1, 3, and 5, and events are merged if they repeat [e.g., original events at $t = 1$ and $t = 2$ at link (A,B) become a single event at $t = 1$].

measures do not fully characterize time-evolving networks; we choose, however, well-established and well-known measures that are of relevance to study epidemic spread and information flow on evolving networks.

The first measure is the burstiness B_s of the link activity [46]. This measure is widely used to characterize temporal patterns on temporal networks. The burstiness depends on the mean m and standard deviation σ of the distribution of same-link interevent times (the interevent time is the time between two subsequent same-link activations) and measures the deviation of the link activity from a Poisson process. Considering the distribution of interevent times of all links collected together, the burstiness is given by

$$B_s = \frac{\sigma - m}{\sigma + m}. \tag{1}$$

The second measure is related to the lifetime L_{ij} (or persistence [47]) of links, that is, the time between the first event, t_{ij}^{first}, and last event, t_{ij}^{last}, on the link (i,j). The link lifetime can be used as a proxy for the real lifetime of contacts [48]. We measure the average lifetime L_s over all K_s links in which $L_{ij} > 0$ (i.e., there are at least two events in the link) to summarize the lifetime of the links in the sampled network, i.e.,

$$L_s = \frac{1}{K_s} \sum_{(i,j)\in E_s, L_{ij}>0} \left(t_{ij}^{\text{last}} - t_{ij}^{\text{first}}\right). \tag{2}$$

The third and fourth measures are related to temporal paths. Temporal paths are particularly relevant in the context of temporal networks because they combine topological and temporal information. They emphasize the role of the timings of events in the connectivity of a node. For example, two nodes may be topologically close (e.g., directly connected by a link) but one may need to wait a long time for this link to be active (i.e., for an interaction event to happen). On the other hand, a more topologically distant pair of nodes (e.g., two links away) may be reached quickly if the interaction events are temporally close. We assume here that, within a time step, a node can only be reached by another node through a direct link. For example, there are no paths connecting nodes A and C if the events (A,B) and (B,C) occur at the same time. An alternative assumption could define a path between A and C in this example [49].

The third measure is the reachability ratio f_s [50]. It is the fraction of pairs of nodes that have at least one temporal path between them and is defined by

$$f_s = \frac{1}{N_s(N_s - 1)} \sum_{i,j=1}^{N_s} \mathbb{1}(\tau_{ij}), \tag{3}$$

where

$$\mathbb{1}(\tau_{ij}) = \begin{cases} 1 & \text{if } \tau_{ij} \text{ exists,} \\ 0 & \text{otherwise.} \end{cases}$$

It can happen that τ_{ij} is finite, whereas τ_{ji} is infinite or vice versa.

The fourth measure is related to the time distance between nodes in the network [34,50,51]. The time distance τ_{ij} is here defined as the time necessary to reach node j from the first appearance (i.e., birth) of node i through the shortest temporal path connecting i and j. If there is no path between nodes i and j, then we set $\tau_{ij} \to \infty$ [51]. We then set

$$\theta_s = \frac{1}{N_s(N_s - 1)} \sum_{i,j=1}^{N_s} \frac{1}{\tau_{ij}} \tag{4}$$

to summarize τ_{ij} over the links. Note that $\tau_{ij} \to \infty$ contributes zero to the sum in Eq. (3) and that both the shortest path from i to j and that from j to i appear in Eq. (3) because τ_{ij} is not equal to τ_{ji} in general. This measure is normalized by $N_s(N_s - 1)$, which gives the total number of possible paths between any two pairs of nodes if all links occur at the same time [49].

For the fifth and sixth measures, we model a susceptible-infected-recovered (SIR) epidemics on the temporal network. In the SIR model, a node can be susceptible (S), infected (I), or recovered (R). Infected nodes can infect susceptible nodes with probability β and recover with probability μ in a time step. For strategy RS, to account for the change in the resolution δ_s (and, consequently, in the contact rate), we rescale the parameters to β/δ_s and μ/δ_s. Rescaling these parameters effectively conserves the contact rate because we assume the events are unweighted; without the rescaling, the infection and recovery probabilities would be overestimated for $\delta_s > \delta$. We start by infecting a single node and leaving all others susceptible. Under the so-called individual-based approximation [37], the dynamics of the probability that node i is infected at time t is given by

$$S_i(t) = S_i(t-1) \prod_{j\in\mathcal{N}_{s\,i}(t)} \phi_j(t), \tag{5}$$

$$I_i(t) = I_i(t-1) + S_i(t-1)\left[1 - \prod_{j\in\mathcal{N}_{s\,i}(t)} \phi_j(t)\right] - \mu I_i(t-1), \tag{6}$$

$$R_i(t) = 1 - S_i(t) - I_i(t), \tag{7}$$

where $\mathcal{N}_{s\,i}(t)$ is the set of neighbors of node i at time t, $\phi_j(t) = 1-(1-\mu)\beta I_j(t-1)$ if there is an event between nodes i and j at time t, and $\phi_j(t) = 1$ otherwise.

We then measure the average number of secondary infections R_s^{eff} and the average final outbreak size Ω_s caused by a single infected node at time 0 for each sampled network [37]. R_s^{eff} is thought to indicate the propensity of an outbreak to become pandemic [52]. The value of Ω_s is not linearly related to R_s^{eff} although a larger Ω_s is expected for larger R_s^{eff} [53]. Under the individual-based approximation, we obtain

$$R_s^{\text{eff}} = \frac{1}{N_s}\sum_{i=1}^{N_s}\left[\sum_{t=1}^{T_s}[1-\phi_i(t)]\sum_{j\in\mathcal{N}_{s\,i}(t)} S_j(t-1)\right] \tag{8}$$

and

$$\Omega_s = \frac{1}{N_s}\sum_{i=1}^{N_s}[I_i(T_s) + R_i(T_s)]. \tag{9}$$

III. RESULTS

A. Network size

Different sampling strategies have a different impact on the number of nodes and events in the sampled networks [Figs. 2(a) and 2(b)]. Reducing the temporal resolution δ_s (strategy RS) has no effect on the number of nodes (N_s) but monotonically decreases the number of events (M_s). This happens because some events repeat at subsequent times. If there is little repetition, reducing the temporal resolution will only slightly decrease the number of events. In the SEX network ($\delta = 1$ day), for example, setting $\delta_s = 63$ days only reduces the number of events by 8.87% [Figs. 2(a) and 2(c)]. This is the reason for the short dashed curve in Fig. 2(a). In contrast, setting $\delta_s = 55$ h in the EMA network ($\delta = 1$ h) reduces the number of events by about 40% [Figs. 2(b) and 2(c)]. The high turnover of nodes (i.e., shorter lifetimes in comparison to the observation time in the original network) in the SEX network explains why the number of nodes falls more substantially in this case than in the EMA network if we reduce T_s (strategy TS). For example, a reduction of about 43% in T_s results in about 37% less nodes in the SEX network [Fig. 2(a)]. For the EMA network, however, the reduction of 48% in T_s implies on only 9.8% less nodes [Fig. 2(b)]. The same reduction in T_s by half results in approximately half the events in both cases [Fig. 2(d)].

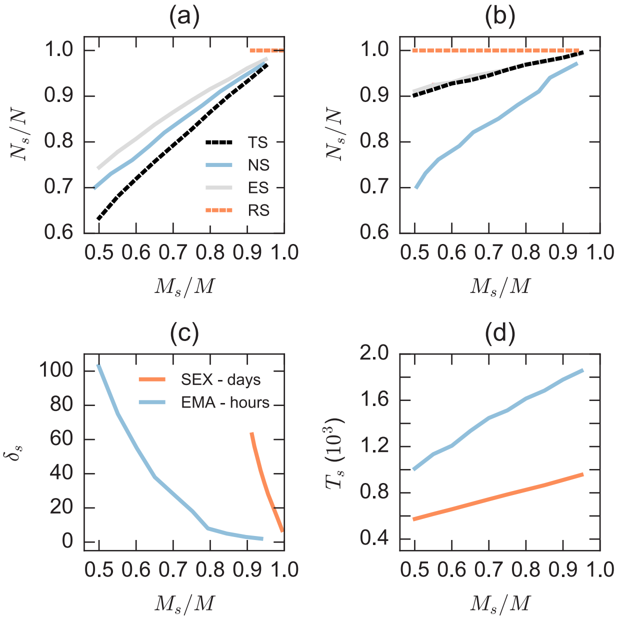

FIG. 2. The fraction of nodes (N_s/N) and events (M_s/M) after sampling the original networks using each of the sampling strategies for the (a) SEX and (b) EMA networks. In (c), we show the fraction of events (M_s/M) for a given temporal resolution (δ_s). In (d), we show the fraction of events (M_s/M) for a given observation time (T_s). Vertical bars for strategies NS and ES correspond to the standard deviation that is only visible if larger than the thickness of the curves.

The uniform sampling of events (strategy ES) has less impact on the number of nodes than the uniform sampling of nodes (strategy NS) if we control for the number of events [Figs. 2(a) and 2(b)]. This happens because a node typically has more than one event with the same or with different neighbors. In strategy ES, highly connected nodes are selected often (proportionally to the number of events [54]) and thus sampled nodes might repeat, decreasing the final number of nodes in the sample. In strategy NS, on the other hand, the selection of nodes brings all their events (to other sampled nodes), implying that fewer nodes are selected (in comparison to strategy ES) for the same number of events.

In the following analyses, we will present the results for COL, FOR, HSC, and GAL using two configurations (A and B) for each strategy. Each configuration corresponds to a fixed number of events M_s. M_s was based on an arbitrarily chosen resolution. That is, we set a resolution δ_s and took the number of events of this sample as reference to be used in the other sampling strategies. For the COL data set, A corresponds to a fraction of 62% ($\delta_s = 48$ h) and B to a fraction of 77% ($\delta_s = 12$ h) of the events of the original network. For the FOR data set, we have, respectively, 56% ($\delta_s = 24$ h) and 74% ($\delta_s = 6$ hours), for HSC, 54% ($\delta_s = 60$ s) and 68% ($\delta_s = 40$ s), and for the GAL data set, we have 57% ($\delta_s = 60$ s) and 70% ($\delta_s = 40$ s).

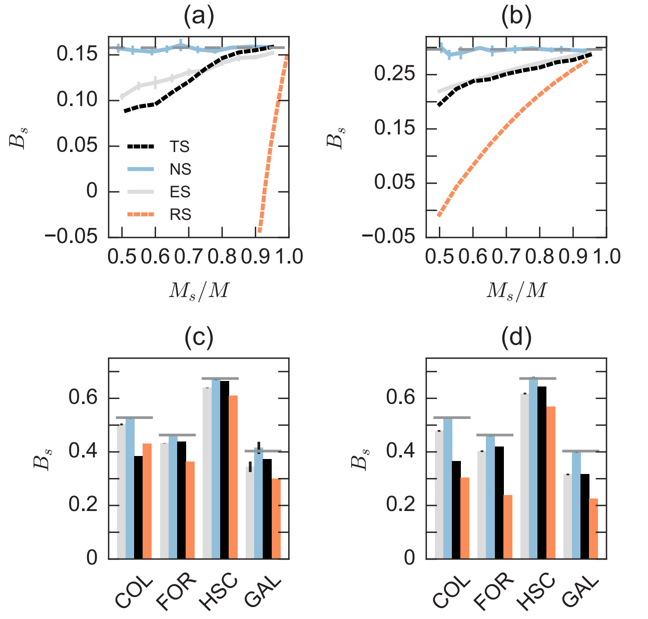

FIG. 3. The burstiness (B_s) of link activity (events) after sampling the original network using each of the sampling strategies for the (a) SEX and (b) EMA networks. The estimation of B_s for the COL, FOR, HSC, and GAL networks with configurations (c) A and (d) B (see Sec. III A). Dashed horizontal lines correspond to the results for the original networks. Vertical bars for strategies NS and ES correspond to the standard deviation that is only visible if larger than the thickness of the curves.

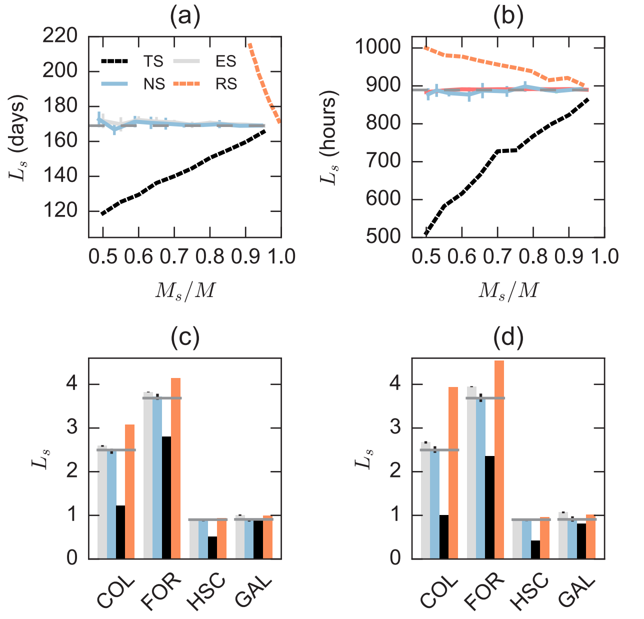

FIG. 4. The average lifetime (L_s) of links after sampling the original network using each of the sampling strategies for the (a) SEX and (b) EMA networks. The estimation of L_s for the COL ($\times 10^2$ h), FOR ($\times 10^2$ d), HSC ($\times 10^4$ s), and GAL ($\times 10^3$ s) networks with configurations (c) A and (d) B. Dashed horizontal lines correspond to the results for the original networks. Vertical bars for strategies NS and ES correspond to the standard deviation.

B. Timings of events

We have found that uniformly sampling nodes (strategy NS) seems to be the best strategy to conserve the burstiness. The value of B_s is robust in both SEX and EMA data sets even when only half of the events are sampled [Figs. 3(a) and 3(b)]. The fact that the number of sampled nodes (by strategy NS) has little impact on the estimation of the burstiness suggests that all nodes follow similar interevent times distributions [i.e., a few nodes are sufficient for an accurate estimation, Figs. 2(a) and 2(b)]. On the other hand, increasing δ_s (strategy RS) has a significant negative effect on B_s. The resolution affects the distribution of interevent times since increasing δ_s filters out short interevent times and reduces the long interevent times, making the signal move towards more regularity (with larger mean and standard deviation). Strategies ES and TS also generate biases, which are considerably smaller than biases given by strategy RS. For different reasons, strategies ES and TS also affect the distribution of interevent times but to a lesser extent than strategy RS. Strategy ES misses a few events and thus increases the average (and standard deviation of the) interevent times. In contrast, strategy TS skips events that could generate long interevent times since the observation time is truncated and thus generates smaller means and standard deviations. Similar results are observed for the other data sets [Figs. 3(c) and 3(d)].

Strategies NS and ES generally give good estimations of the average lifetime of links L_s for all data sets [Figs. 4(a)–4(d)]. The uniform sampling of events or nodes decreases the lifetime of some links but also sometimes does not sample any event of a particular link (i.e., some links and nodes may not be sampled at all). The smaller K_s possibly compensates the decrease in the lifetimes such that the average L_s is little affected. Strategy TS introduces cutoffs on the lifetimes of both links and nodes since sampling is limited within the observation time $[0,T_s]$. Consequently, the lifetime is underestimated. The case of GAL is special because visitors explore the museum in groups at allocated times, meaning that links form and disappear before T_s [Figs. 4(c) and 4(d)]. Finally, strategy RS tends to overestimate L_s because increasing δ_s is equivalent to rounding down the times of births and deaths. The rounding down leads to an overall increase in the lifetime of links and a decrease in K_s since links with a single event are not included in the average.

C. Temporal paths

The reachability, f_s, changes substantially for the SEX and HSC networks but not as much for the other networks [Figs. 5(a)–5(d)]. For example, in the original SEX network about 34% of the pairs of nodes were reachable in contrast to about 94% in the EMA original network. After sampling, only strategy RS decreases f_s in the EMA network. However, the difference with the original value is small, e.g., 6.4% in the sampled EMA network containing about 50% of the original events. This is considerably less than in the case of the SEX network that shows a difference of 55.9% to the original value for the same strategy RS [Figs. 5(a) and 5(b)]. The generally observed low biases generated by strategies NS and ES result from the redundancy of paths, i.e., the fact that there are multiple paths connecting the same pairs of nodes at distinct times. The absence of some events thus has little impact on f_s. The same redundancy is also observed for example in the

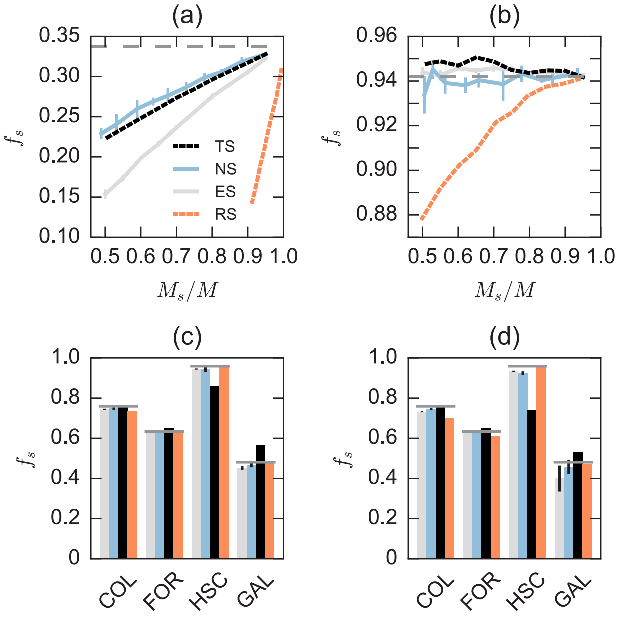

FIG. 5. The fraction of temporal paths (f_s) after sampling the original network using each of the sampling strategies for the (a) SEX and (b) EMA networks. The estimation of f_s for the COL, FOR, HSC, and GAL networks with configurations (c) A and (d) B. Dashed horizontal lines correspond to the results for the original networks. Vertical bars for strategies NS and ES correspond to the standard deviation.

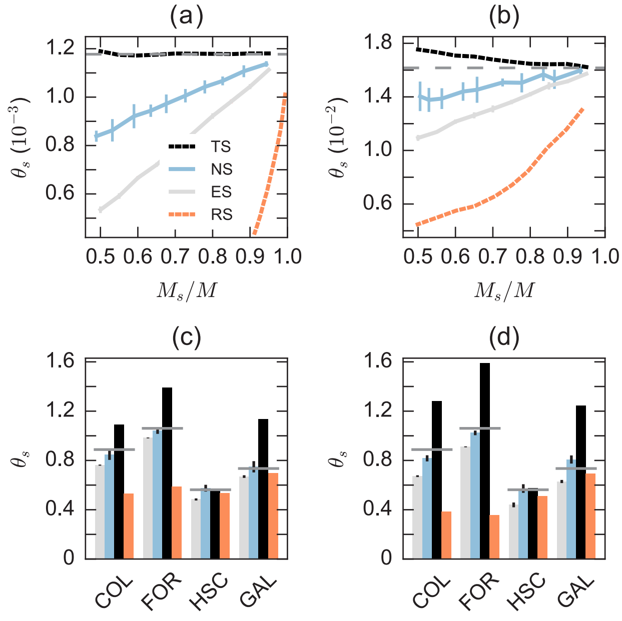

FIG. 6. The average of the inverse of the temporal distance (θ_s) between any pair of nodes after sampling the original network using each of the sampling strategies for the (a) SEX and (b) EMA networks. The estimation of θ_s for the COL ($\times 10^2$ h), FOR ($\times 10^2$ d), HSC ($\times 10^3$ s), and GAL ($\times 10^3$ s) networks with configurations (c) A and (d) B. Dashed horizontal lines correspond to the results for the original networks. Vertical bars for strategies NS and ES correspond to the standard deviation.

SEX network but at a lesser extent, possibly because of the relatively smaller density of events in the SEX network in comparison to the EMA network (see Table I). Furthermore, the low observed biases of strategy TS (for most data sets) indicate that the number of existing shortest paths decreases at the same rate as the number of potential paths $[N_s(N_s-1)]$, for smaller T_s. The biases observed for SEX and HSC data sets, on the other hand, thus indicate that the new sampled nodes (introduced in the sample for increasing T_s) do not result in the same number of new paths as the number of potential paths that could exist (i.e., f_s decreases with increasing T_s).

Figures 6(a)–6(d) shows that the statistics of the duration of the temporal paths between nodes, θ_s, changes for EMA, COL, FOR, and GAL for strategy TS. For the SEX and HSC data sets, this strategy generates very low biases. Although several shortest temporal paths are formed before T_s, some only exist if we increase T_s. Therefore, if we truncate the data to T_s, then the summation term in θ_s may decrease. But since nodes are also removed (i.e., lower N_s), the overall value of θ_s increases. For the SEX and HSC data sets, the decrease in the summation term is equivalent to the decrease in the number of potential shortest paths $[N_s(N_s-1)]$. On the other hand, strategy RS results in considerably different values for the SEX, EMA, COL, and FOR data sets. Strategy RS generates larger biases than the other strategies because higher δ_s rounds down the timings of events, collapsing many links to the same time interval and thus removing several temporal paths between nodes, that in turn results in smaller θ_s. Remember that in our definition, only directly connected nodes have a temporal path within the same time step. For the other two strategies (NS and ES), uniform sampling of nodes or events increases, on average, the temporal distances between nodes. The higher θ_s given by strategy NS, in comparison to strategy ES, is possibly a result of a smaller N_s obtained by strategy NS in comparison to the N_s obtained by strategy ES (see Fig. 2 for the SEX and EMA data sets). The relatively smaller biases in the EMA data set in comparison to the SEX data set are likely a result of higher redundancy of paths in the EMA network, as discussed in the previous paragraph.

D. Epidemic variables

We set $\beta = 0.5$ and $\mu = 0.001$ to simulate a stochastic epidemic process. These values were chosen because they generate relatively large epidemic outbreaks in all original networks, and thus facilitate the understanding and discussion of the mechanisms regulating the epidemic process.

We first look at the average number of secondary infections, R_s^{eff}. Strategy TS results in a relatively small increase in R_s^{eff} for most data sets, whereas strategies NS and ES result in a small decrease for all data sets [Figs. 7(a)–7(d)]. The estimations of R_s^{eff} given by the sampled networks indicate that the systems remain above the epidemic threshold of $R_s^{\text{eff}} = 1$ for this particular set of parameters and that an epidemic outbreak will likely occur. Since the value of R_s^{eff} also indicates how difficult is to avoid an epidemic outbreak, the estimations given by the sampled networks generally suggest that an outbreak might be easier to control than indicated by the original network (i.e., R_s^{eff} is closer to one in the sampled networks). The results for strategy RS are substantially far from the value given by the original network for the SEX, EMA, and COL data sets

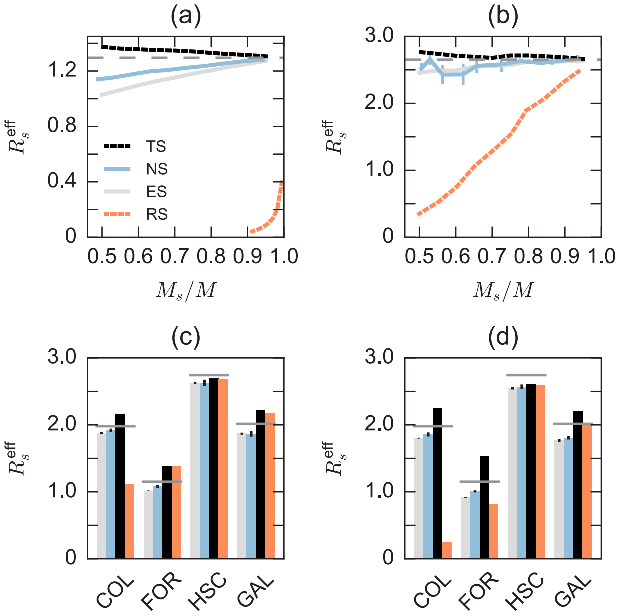

FIG. 7. The average number of secondary infections (R_s^{eff}) after sampling the original network using each of the sampling strategies for the (a) SEX and (b) EMA networks. The estimation of R_s^{eff} for the COL, FOR, HSC, and GAL networks with configurations (c) A and (d) B. Dashed horizontal lines correspond to the results for the original networks. Vertical bars for strategies NS and ES correspond to the standard deviation.

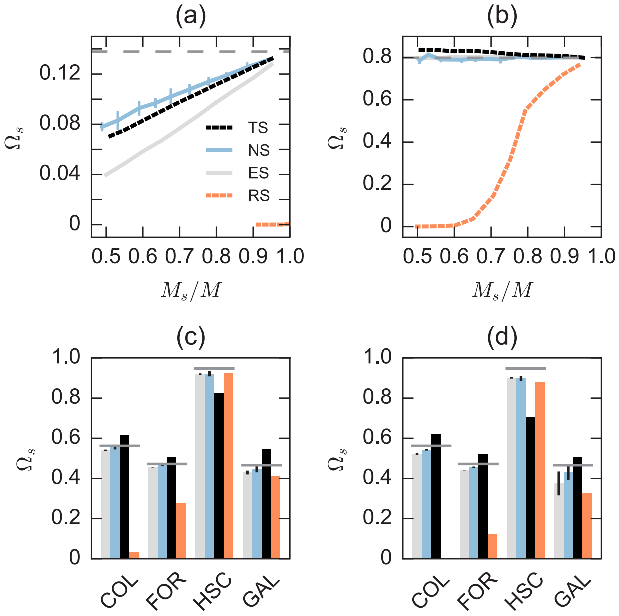

FIG. 8. The average outbreak size (Ω_s) after sampling the original network using each of the sampling strategies for the (a) SEX and (b) EMA networks. The estimation of Ω_s for the COL, FOR, HSC, and GAL networks with configurations (c) A and (d) B. Dashed horizontal lines correspond to the results for the original networks. Vertical bars for strategies NS and ES correspond to the standard deviation.

but not for the FOR, HSC, and GAL data sets. The low biases produced by strategies NS and ES across the different data sets are explained by the fact that the infection process is temporally finite. Many events do not actually contribute to the spread of the infection given the stochastic nature of the process, i.e., the absence of randomly selected interaction events has a relatively little importance to avoid infection events. The negative effect of the absence of interaction events is lower for the EMA network in which events repeat more often than in the SEX network. Therefore, the same neighbor has more chances of being infected in subsequent times in EMA than in the SEX network. This is related to the results observed for θ_s (Fig. 5) and f_s (Fig. 6), where a substantial absence of events generated small biases for most networks. Strategy TS also performs well because of the finite time of the infection period that makes most infection events occur before T_s. If the infection period is long (small μ) or the infection probability is small, the biases given by strategy TS are expected to be larger. Since the number of nodes is smaller in comparison to the original networks, R_s^{eff} becomes slightly overestimated by strategy TS. On the other hand, strategy RS generates large biases. Increasing δ_s alters the infection potential through a particular event and extends the infection period because of the rescaling of the infection and recovery probabilities, respectively. For example, in the SEX data set, if $\delta_s = 7$ days, then the effective infection probability is $\beta_s = \beta/7 \sim 0.07$; this infection probability is too low. Combined with the fact that the number of events (of a single node to different neighbors) at a given time step does not increase much for increasing δ_s, very few neighbors may be infected by an infectious node [Fig. 7(a)]. In the EMA network, on the other hand, there will be more events (connecting different nodes) at a single time step and thus there is a higher chance of infecting some neighbors. See also Fig. 2(c) for the correspondence between M_s/M and δ_s for the SEX and EMA data sets.

Figures 8(a)–8(d) shows that the final outbreak size, Ω_s, is close to zero for strategy RS applied to the SEX network, to the EMA network when approximately 65% (or less) of the events are sampled, and to the COL network. For the other three sampling strategies, Ω_s is similar between the sampled and original networks for most data sets but increasingly different for smaller samples in the case of the SEX network. This is again explained by the fact that events repeat over time (less often in the SEX network). This repetition of events creates redundancies of temporal paths. In the absence of several events (by any of these three strategies), various potential infection routes remain between the nodes, and the epidemic may still grow. The biases should increase for smaller infection probabilities since an infection event will be less likely through a particular interaction event.

IV. CONCLUSIONS

Our analyses indicate that generally both measures related to link activity are little affected by uniform sampling of nodes. This strategy also had very good performance for estimation of the statistics of temporal paths and epidemics for all network data sets but the sexual contacts data set. These results likely explain the high performance of recently proposed methods to reconstruct temporal networks [55,56]. That is, the temporal patterns extracted from a small sample of the temporal network are sufficient to generate larger temporal networks with realistic temporal properties. However, more research is necessary to validate these methods on diverse types of networks. Uniform sampling of events have also performed well for most statistics on most data sets. Although less efficient than uniform sampling of nodes, sampling of events may be an option when continuously collecting network data. For example, for a given number of nodes, at each time step a fraction of links may be selected and stored as time evolves ("on-line sampling"). This procedure is expected to produce better samples than truncating the observation time. In fact, truncating the observation time produced mixed results. For some networks, this sampling strategy did not affect much the statistics but for some other data sets, relatively high biases are observed (e.g., for lifetime and for the temporal distance). Although performing well in some cases, the poorest performance was obtained when varying the temporal resolution. In some networks, there are many repetitions of events. Therefore, merging the events on the same link by reducing the temporal resolution implies small changes in the temporal network structure. On the other hand, if there are few repetitions of events, then the network might look substantially different at each temporal resolution, consequently affecting the statistics. Using a different methodology, previous research suggests that for a set of epidemiological parameters a high temporal resolution might not be necessary to study simulated epidemics in some systems [15]. However, to different extents, the temporal resolution seems to affect the dynamics of consensus and random walks in some systems [22,26]. The statistics used in our study are extensive but do not capture all the patterns observed in temporal networks. Recent research has pointed out, for example, the importance of the correlations in the ordering of events and how such correlations affect path-based centrality measures in time-evolving networks [57]. Methods based on Markov chains and high-order networks have been also proposed to capture multiple length scales simultaneously and to identify the optimal number of layers in those evolving networks [58].

In general, we have identified differences in the magnitude of the biases on various statistics and real-life networks. Given our results, we advice to avoid reducing much the temporal resolution but instead, if possible, we recommend uniform sampling of nodes to conserve several of the properties of temporal networks. Given the complexity of temporal structures and particularities of data sets, there is, however, no guarantee that temporal-structural patterns will be conserved. The choice of a sampling strategy thus depends on which network property one wishes to study and should be considered for the specific problem, context, and goals of the research. In general, there is room for sampling design. In practice, it is likely to combine all proposed sampling strategies in a data collection project. It is difficult to predict the consequences of combining them since positive bias by one strategy may compensate negative bias by another strategy or biases may simply add up or multiply. Nevertheless, our study of the effects of separately applying each sampling strategy will likely improve data collection by helping researchers to make informed decisions and question the limitations of their own data sets.

[1] M. E. J. Newman, *Networks: An Introduction* (Oxford University Press, Oxford, 2010).

[2] L. F. Costa, O. N. Oliveira Jr., G. Travieso, F. A. Rodrigues, P. R. Villas Boas, L. Antiqueira, M. P. Viana, and L. E. C. da Rocha, Analyzing and modeling real-world phenomena with complex networks: A survey of applications, Adv. Phys. **60**, 329 (2011).

[3] S. L. Lohr, *Sampling: Design and Analysis* (Cengage Learning, Boston, MA, 2009).

[4] S. Sudman, M. G. Sirken, and C. D. Cowan, Sampling rare and elusive populations, Science **240**, 991 (1988).

[5] D. Achlioptas, A. Clauset, D. Kempe, and C. Moore, On the bias of traceroute sampling: Or, power-law degree distributions in regular graphs, J. ACM **56**, 21 (2009).

[6] S. H. Lee, P.-J. Kim, and H. Jeong, Statistical properties of sampled networks, Phys. Rev. E **73**, 016102 (2006).

[7] M. A. Serrano, M. Boguna, and A. Vespignani, Extracting the multiscale backbone of complex weighted networks, Proc. Nat. Acad. Sci. U.S.A. **106**, 6483 (2009).

[8] P. Holme, Modern temporal network theory: A colloquium, Eur. Phys. J. B **88**, 234 (2015).

[9] J.-P. Eckmann, E. Moses, and D. Sergi, Entropy of dialogues creates coherent structures in e-mail traffic, Proc. Natl. Acad. Sci. U.S.A. **101**, 14333 (2004).

[10] N. Masuda and P. Holme, Predicting and controlling infectious disease epidemics using temporal networks, F1000Prime Rep. **5**, 6 (2013).

[11] L. E. C. Rocha, Dynamics of air transport networks: A review from a complex systems perspective, Chin. J. Aeronaut. **30**, 469 (2017).

[12] C. S. Riolo, J. S. Koopman, and J. S. Chick, Methods and measures for the description of epidemiological contact networks, J. Urban Health **78**, 446 (2001).

[13] N. H. Fefferman and K. L. Ng, How disease models in static networks can fail to approximate disease in dynamic networks, Phys. Rev. E **76**, 031919 (2007).

[14] L. E. C. Rocha, F. Liljeros, and P. Holme, Simulated epidemics in an empirical spatiotemporal network of 50, 185 sexual contacts, PLoS Comput. Biol. **7**, e1001109 (2011).

[15] 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 a SEIR infectious disease model on the dynamic contact network of conference attendees, BMC Med. **9**, 87 (2011).

[16] R. Lambiotte, L. Tabourier, and J.-C. Delvenne, Burstiness and spreading on temporal networks, Eur. Phys. J. B **86**, 320 (2013).

[17] L. Speidel, K. Klemm, V. M. Eguíluz, and N. Masuda, Temporal interactions facilitate endemicity in the susceptible-infected-susceptible epidemic model, New J. Phys. **18**, 073013 (2016).

[18] L. Lamport, Time, clocks, and the ordering of events in a distributed system, Commun. ACM **21**, 558 (1978).

[19] J. Moody, The importance of relationship timing for diffusion, Soc. Forces **81**, 25 (2002).

[20] A. Vazquez, B. Rácz, A. Lukács, and A.-L. Barabási, Impact of Non-Poissonian Activity Patterns on Spreading Processes, Phys. Rev. Lett. **98**, 158702 (2007).

[21] 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).

[22] B. Ribeiro, N. Perra, and A. Baronchelli, Quantifying the effect of temporal resolution on time-varying networks, Sci. Rep. **3**, 1 (2013).

[23] I. Scholtes, N. Wider, R. Pfitzner, A. Garas, C. J. Tessone, and F. Schweitzer, Causality-driven slow-down and speed-up of diffusion in non-Markovian temporal networks, Nat. Commun. **5**, 5024 (2014).

[24] J.-C. Delvenne, R. Lambiotte, and L. E. C. Rocha, Diffusion on networked systems is a question of time or structure, Nat. Commun. **6**, 7366 (2015).

[25] N. Masuda, K. Klemm, and V. M. Eguíluz, Temporal Networks: Slowing Down Diffusion by Long Lasting Interactions, Phys. Rev. Lett. **111**, 188701 (2013).

[26] A. Cardillo, G. Petri, V. Nicosia, R. Sinatra, J. Gómez-Gardeñes, and V. Latora, Evolutionary dynamics of time-resolved social interactions, Phys. Rev. E **90**, 052825 (2014).

[27] N. Masuda, Accelerating coordination in temporal networks by engineering the link order, Sci. Rep. **6**, 22105 (2016).

[28] P. Holme and L. E. C. Rocha, Sensitivity to temporal and topological misinformation in predictions of epidemic outbreaks, in *Temporal Network Epidemiology, Theoretical Biology*, edited by N. Masuda and P. Holme (Springer, Singapore, 2017).

[29] P. Holme and L. E. C. Rocha, Impact of misinformation in temporal network epidemiology, arXiv:1704.02406.

[30] G. Krings, M. Karsai, S. Bernhardsson, V. D. Blondel, and J. Saramäki, Effects of time window size and placement on the structure of an aggregated communication network, Eur. Phys. J. Data Sci. **1**, 4 (2012).

[31] A. Barrat, C. Cattuto, V. Colizza, F. Gesualdo, L. Isella, E. Pandolfi, J.-F. Pinton, L. Rava, C. Rizzo, M. Romano, J. Stehlé, A. E. Tozzi, and W. Van den Broeck, Empirical temporal networks of face-to-face human interactions, Eur. Phys. J. Spec. Top. **222**, 1295 (2013).

[32] A. Barrat and C. Cattuto, Temporal networks of face-to-face human interactions, in *Temporal Networks*, edited by P. Holme and J. Saramäki (Springer, Berlin, 2013), pp. 191–216.

[33] 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).

[34] R. K. Pan and J. Saramäki, Path lengths, correlations, and centrality in temporal networks, Phys. Rev. E **84**, 016105 (2011).

[35] 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).

[36] C. L. Vestergaard and M. Génois, Temporal Gillespie algorithm: Fast simulation of contagion processes on time-varying networks, PLoS Comput. Biol. **11**, e1004579 (2015).

[37] L. E. C. Rocha and N. Masuda, Individual-based approach to epidemic processes on arbitrary dynamic contact networks, Sci. Rep. **6**, 1 (2016).

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

[39] J. Saramäki and E. Moro, From seconds to months: An overview of multi-scale dynamics of mobile telephone calls, Eur. Phys. J. B **88**, 164 (2015).

[40] L. E. C. Rocha, F. Liljeros, and P. Holme, Information dynamics shape the sexual networks of Internet-mediated prostitution, Proc. Natl. Acad. Sci. U.S.A. **107**, 5706 (2010).

[41] F. Karimi, V. C. Ramenzoni, and P. Holme, Structural differences between open and direct communication in an online community, Physica A **414**, 263 (2014).

[42] P. Panzarasa, T. Opsahl, and K. M. Carley, Patterns and dynamics of users' behavior and interaction: Network analysis of an online community, J. Am. Soc. Inf. Sci. Technol. **60**, 911 (2009).

[43] 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).

[44] W. van den Broeck, M. Quaggiotto, L. Isella, A. Barrat, and C. Cattuto, The making of sixty-nine days of close encounters at the science gallery, Leonardo **45**, 285 (2012).

[45] D. N. Politis, J. P. Romano, and M. Wolf, *Subsampling* (Springer, New York, 1999).

[46] K.-I. Goh and A.-L. Barabasi, Burstiness and memory in complex systems, Europhys. Lett. **81**, 48002 (2008).

[47] A. Clauset and N. Eagle, Persistence and periodicity in a dynamic proximity network, in *DIMACS Workshop on Computational Methods for Dynamic Interaction Networks* (2007), pp. 1–5.

[48] P. Holme, Network dynamics of ongoing social relationships, Europhys. Lett. **64**, 427 (2003).

[49] J. Tang, M. Musolesi, C. Mascolo, and V. Latora, Characterising temporal distance and reachability in mobile and online social networks, ACM SIGCOMM Comput. Commun. Rev. **40**, 118 (2010).

[50] P. Holme, Network reachability of real-world contact sequences, Phys. Rev. E **71**, 046119 (2005).

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

[52] D. Dietz, The estimation of the basic reproduction number for infectious diseases, Stat. Methods Med. Res. **2**, 23 (1993).

[53] P. Holme and N. Masuda, The basic reproduction number as a predictor for epidemic outbreaks in temporal networks, PLoS One **10**, e0120567 (2015).

[54] S. L. Feld, Why your friends have more friends than you do, Am. J. Sociol. **96**, 1464 (1991).

[55] 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).

[56] 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).

[57] I. Scholtes, N. Wider, and A. Garas, Higher-order aggregate networks in the analysis of temporal networks: Path structures and centralities, Eur. Phys. J. B **89**, 61 (2016).

[58] I. Scholtes, When is a network a network? Multi-order graphical model selection in pathways and temporal networks, in *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (ACM, New York, NY, 2017), pp. 1037–1046.