Objective measures for sentinel surveillance in network epidemiology

Petter Holme[*]

Institute of Innovative Research, Tokyo Institute of Technology, Nagatsuta-cho 4259, Midori-ku, Yokohama, Kanagawa 226-8503, Japan

Assume one has the capability of determining whether a node in a network is infectious or not by probing it. Then problem of optimizing sentinel surveillance in networks is to identify the nodes to probe such that an emerging disease outbreak can be discovered early or reliably. Whether the emphasis should be on early or reliable detection depends on the scenario in question. We investigate three objective measures from the literature quantifying the performance of nodes in sentinel surveillance: the time to detection or extinction, the time to detection, and the frequency of detection. As a basis for the comparison, we use the susceptible-infectious-recovered model on static and temporal networks of human contacts. We show that, for some regions of parameter space, the three objective measures can rank the nodes very differently. This means sentinel surveillance is a class of problems, and solutions need to chose an objective measure for the particular scenario in question. As opposed to other problems in network epidemiology, we draw similar conclusions from the static and temporal networks. Furthermore, we do not find one type of network structure that predicts the objective measures, i.e., that depends both on the data set and the SIR parameter values.

I. INTRODUCTION

Infectious diseases are a big burden to public health. Their epidemiology is a topic wherein the gap between the medical and theoretical sciences is not so large. Several concepts of mathematical epidemiology—like the basic reproductive number or core groups [1–3]—have entered the vocabulary of medical scientists. Traditionally, authors have modeled disease outbreaks in society by assuming any person to have the same chance of meeting anyone else at any time. This is of course not realistic, and improving this point is the motivation for network epidemiology: epidemic simulations between people connected by a network [4]. One can continue increasing the realism in the contact patterns by observing that the timing of contacts can also have structures capable of affecting the disease. Studying epidemics on time-varying contact structures is the basis of the emerging field of temporal network epidemiology [5–8].

One of the most important questions in infectious disease epidemiology is to identify people, or in more general terms, units, that would get infected early and with high likelihood in an infectious outbreak. This is the *sentinel surveillance* problem [9,10]. It is the aspect of node importance, which is the one most actively used in public health practice. Typically, it works by selecting some hospitals (clinics, cattle farms, etc.) to screen, or more frequently test, for a specific infection [11].

Defining an objective measure—a quantity to be maximized or minimized—for sentinel surveillance is not trivial. It depends on the particular scenario one considers and the means of interventions at hand. If the goal for society is to detect as many outbreaks as possible, it makes sense to choose sentinels to maximize the fraction of detected outbreaks [9]. If the objective rather is to discover outbreaks early, then one could choose sentinels that, if infected, are infected early [10,12]. Finally, if the objective is to stop the disease as early as possible, it makes sense to measure the time to extinction or detection (infection of a sentinel) [13]. See Fig. 1 for an illustration. To restrict ourselves, we will focus on the case of one sentinel. If one has more than one sentinel, the optimal set will most likely not be the top nodes of a ranking according to the three measures above. Their relative positions in the network also matter (they should not be too close to each other) [13].

In this paper, we study and characterize our three objective measures. We base our analysis on 38 empirical data sets of contacts between people. We analyze them both in temporal and static networks. The reason we use empirical contact data, rather than generative models, as the basis of this study is twofold. First, there are so many possible structures and correlations in temporal networks that one cannot tune them all in models [8]. It is also hard to identify the most important structures for a specific spreading phenomenon [8]. Second, studying empirical networks makes this paper—in addition to elucidating the objective measures of sentinel surveillance—a study of human interaction. We can classify data sets with respect how the epidemic dynamics propagate on them. As mentioned above, in practical sentinel surveillance, the network in question is rather one of hospitals, clinics or farms. One can, however, also think of sentinel surveillance of individuals, where high-risk individuals would be tested extra often for some diseases.

In the remainder of the paper, we will describe the objective measures, the structural measures we use for the analysis, and the data sets, and we will present the analysis itself. We will primarily focus on the relation between the measures, secondarily on the structural explanations of our observations.
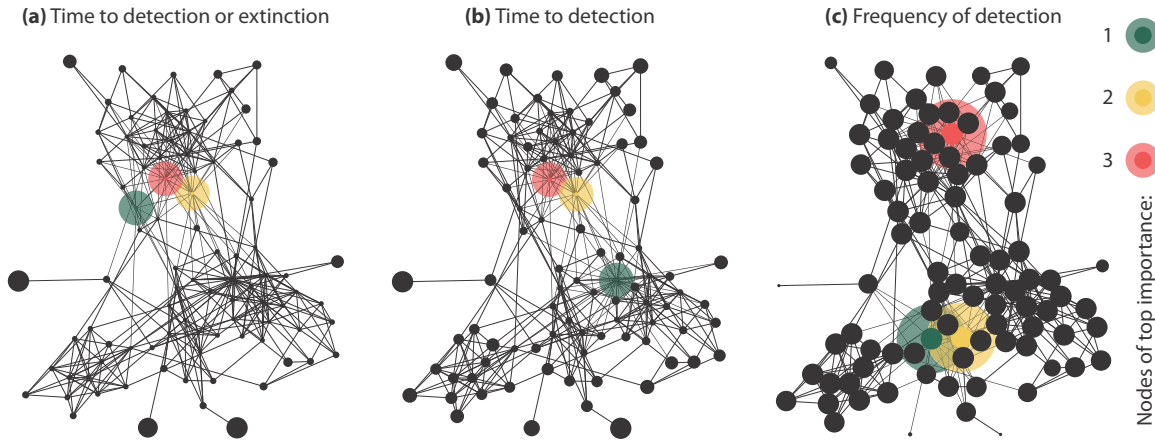
---

*holme@cns.pi.titech.ac.jp

FIG. 1. Objective measures for the *Office* data with the infection rate $\beta = 1$: (a) the time to detection or extinction, (b) time to detection, (c) frequency of detection. The area of the circles are proportional to the respective objective measure. This means that in panel (c) the most important node is the largest, while in (a) and (b) it is the smallest. The three most important nodes of each panel are highlighted.

II. METHODS

A. Objective measures

1. Time to detection or extinction

Assume that the objective of society is to end outbreaks as soon as possible. If an outbreak dies by itself, that is fine. Otherwise, one would like to detect it so it could be mitigated by interventions. In this scenario, a sensible objective measure would be the time for a disease to either go extinct or be detected by a sentinel: *the time to detection or extinction* t_x [13].

2. Time to detection

Suppose that, in contrast to the situation above, the priority is not to save society from the epidemics as soon as possible, but just to detect outbreaks fast. This could be the case if one would want to get a chance to isolate a pathogen, or start producing a vaccine, as early as possible, maybe to prevent future outbreaks of the same pathogen at the earliest possibility. Then one would seek to minimize the time for the outbreak to be detected conditioned on the fact that it is detected: *the time to detection* t_d.

3. Frequency of detection

For the time to detection, it does not matter how likely it is for an outbreak to reach a sentinel. If the objective is to detect as many outbreaks as possible, the corresponding measure should be the expected frequency of outbreaks to reach a node: the *frequency of detection* f_d.

Note that for this measure a large value means the node is a good sentinel, whereas for t_x and t_d a good sentinel has a low value. This means that when we correlate the measures, a similar ranking between t_x and f_d or t_d and f_d yields a negative correlation coefficient. Instead of considering the inverse times, or similar, we keep this feature and urge the reader to keep this in mind.

B. Reducing temporal networks to static networks

There are many possible ways to reduce our empirical temporal networks to static networks. The simplest method would be to just include a link between any pair of nodes

that has at least one contact during the course of the data set. This would however make some of the networks so dense that the static network structure of the node-pairs most actively in contact would be obscured. For our purpose, we primarily want our network to span many types of network structures that can impact epidemics. Without any additional knowledge about the epidemics, the best option is to threshold the weighted graph where an edge (i, j) means that i and j had more than θ contacts in the data set. In this work, we assume that we do not know what the per-contact transmission probability β is (this would anyway depend on both the disease and precise details of the interaction). Rather we scan through a very large range of β values. Since we anyway to that, there is no need either to base the choice of θ on some epidemiological argument, or to rescale β after the thresholding. Note that the rescaled β would be a non-linear function of the number of contacts between i and j. (Assuming no recovery, for an isolated link with ν contacts, the transmission probability is $1 - (1 - \beta)^{\nu}$.) For our purpose the only thing we need is that the rescaled β is a monotonous function of β for the temporal network (which is true). To follow a simple principle, we omit all links with a weight less than the median weight θ.

C. Disease simulations

We simulate disease spreading by the SIR dynamics, the canonical model for diseases that gives immunity upon recovery [2,14]. For static networks, we use the standard Markovian version of the SIR model [15]. That is, we assume that diseases spread over links between susceptible and infectious nodes the infinitesimal time interval dt with a probability $\beta \, dt$. Then, an infectious node recovers after a time that is exponentially distributed with average $1/\nu$. The parameters β and ν are called *infection rate* and *recovery rate*, respectively. We can, without loss of generality, put $\nu = 1/T$ (where T is the duration of the sampling). For other ν values, the ranking of the nodes would be the same (but the values of the t_x and t_d would be rescaled by a factor ν). We will scan an exponentially increasing progression of 200 values of β, from 10^{-3} to 10. The code for the disease simulations can be downloaded [16].

For the temporal networks, we use a definition as close as possible to the one above. We assume an exponentially distributed duration of the infectious state with mean $1/\nu$. We assume a contact between an infectious and susceptible node results in a new infection with probability β. In the case of temporal networks, one cannot reduce the problem to one parameter. Like for static networks, we sample the parameter values in exponential sequences in the intervals $0.01 \leqslant \beta \leqslant 1$ and $0.01 \leqslant \nu/T \leqslant 1$ respectively. For temporal networks, with our interpretation of a contact, $\beta > 1$ makes no sense, which explains the upper limit. Furthermore, since temporal networks usually are effectively sparser (in terms of the number of possible infection events per time), the smallest β values will give similar results, which is the reason for the higher cutoff in this case.

For both temporal and static networks, we assume the outbreak starts at one randomly chosen node. Analogously, in the temporal case we assume the disease is introduced with equal probability at any time throughout the sampling period. For every data set and set of parameter values, we sample 10^7 runs of epidemic simulations.

D. Empirical networks

As motivated in the Introduction, we base our study on empirical temporal networks. All networks that we study record contacts between people and falls into two classes: human proximity networks and communication networks. Proximity networks are, of course, most relevant for epidemic studies, but communication networks can serve as a reference (and it is interesting to see how general results are over the two classes). The data sets consist of anonymized lists of two identification numbers in contact and the time since the beginning of the contact.

Many of the proximity data sets we use come from the Sociopatterns project [17]. These data sets were gathered by people wearing radio-frequency identification (RFID) sensors that detect proximity between 1 and 1.5 m. One such datasets comes from a conference, Hypertext 2009, (*Conference 1*) [18], another two from a primary school (*Primary School*) [19] and five from a high school (*High School*) [20], a third from a hospital (*Hospital*) [21], a fourth set of five data sets from an art gallery (*Gallery*) [22], a fifth from a workplace (*Office*) [23], and a sixth from members of five families in rural *Kenya* [24]. The *Gallery* data sets consist of several days where we use the first five.

In addition to data gathered by RFID sensors, we also use data from the longer-range (around 10m) Bluetooth channel. The *Cambridge 1* [25] and *2* [26] datasets were measured by the Bluetooth channel of sensors (iMotes) worn by people in and around Cambridge, UK. *St Andrews* [27], *Conference 2* [25], and *Intel* [25] are similar data sets tracing contacts at, respectively, the University of St. Andrews, the conference Infocom 2006, and the Intel research laboratory in Cambridge, UK. The *Reality* [28] and *Copenhagen Bluetooth* [29] data sets also come from Bluetooth data, but from smartphones carried by university students. In the *Romania* data, the WiFi channel of smartphones was used to log the proximity between university students [30], whereas the *WiFi* dataset links students of a Chinese university that are logged onto the same WiFi router.

For the *Diary* data set, a group of colleagues and their family members were self-recording their contacts [31]. Our final proximity data, the *Prostitution* network, comes from from self-reported sexual contacts between female sex workers and their male sex buyers [32]. This is a special form of proximity network since contacts represent more than just proximity.

Among the data sets from electronic communication, *Facebook* comes from the wall posts at the social media platform Facebook [33]. *College* is based on communication at a Facebook-like service [34]. *Dating* shows interactions at an early Internet dating website [35]. *Messages* and *Forum* are similar records of interaction at a film community [36]. *Copenhagen Calls* and *Copenhagen SMS* consist of phone calls and text messages gathered in the same experiment as *Copenhagen Bluetooth* [29]. Finally, we use four data sets of e-mail communication. One, *E-mail 1*, recorded all e-mails to and from a group of accounts [37]. The other three, *E-mail 2* [38], *3* [39], and *4* [40] recorded e-mails within a set of accounts.

We list basic statistics—sizes, sampling durations, etc.—of all the data sets in Table I.

E. Static network descriptors

To gain further insight into the network structures promoting the objective measures, we correlate the objective measures with quantities describing the position of a node in the static networks. Since many of our networks are fragmented into components, we restrict ourselves to measures that are well defined for disconnected networks. Otherwise, in our selection, we strive to cover as many different aspects of node importance as we can.

1. Degree

Degree is simply the number of neighbors of a node. It usually presented as the simplest measure of centrality and one of the most discussed structural predictors of importance with respect to disease spreading [42]. (Centrality is a class of measures of a node's position in a network that try to capture what a "central" node is; i.e., ultimately centrality is not more well-defined than the vernacular word.) It is also a local measure in the sense that a node is able to estimate its degree, which could be practical when evaluating sentinel surveillance in real networks.

2. Subgraph centrality

Subgraph centrality is based on the number of closed walks a node is a member of. (A walk is a path that could be overlapping itself.) The number of paths from node i to itself is given by $\mathbf{A}_{ii}^{\lambda}$, where $\mathbf{A}$ is the adjacency matrix and λ is the length of the path. Reference [43] argues that the best way to weigh paths of different lengths together is through the formula

$$C_S(i) = \sum_\lambda \frac{\mathbf{A}_{ii}^\lambda}{\lambda!}. \tag{1}$$

3. Component size

As mentioned, several of the data sets are fragmented (even though the largest connected component dominates components of other sizes). In the limit of high transmission

TABLE I. Basic statistics of the empirical temporal networks. N is the number of nodes, C is the number of contacts, T is the total sampling time, Δt is the time resolution of the data set, M is the number of links in the projected and thresholded static networks, and θ is the threshold.

Data set	N	C	T	Δt	M	θ	Ref.
Conference 1	113	20 818	2.50 d	20 s	1 321	2	[18]
Conference 2	198	327 333	2.95 d	20 s	775	75	[25]
Hospital	75	32 424	4.02 d	20 s	582	8	[21]
Reality	63	26 260	8.63 h	5 s	421	3	[28]
Office	92	9 827	11.4 d	20 s	389	3	[23]
Primary School 1	236	60 623	8.64 h	20 s	299	3	[19]
Primary School 2	238	65 150	8.58 h	20 s	257	3	[19]
Romania	42	1 748 401	62.8 d	1 m	128	61	[30]
High School 1	312	28 780	4.99 h	20 s	1 385	2	[20]
High School 2	310	47 338	8.99 h	20 s	1 601	2	[20]
High School 3	303	40 174	8.99 h	20 s	1 096	3	[20]
High School 4	295	37 279	8.99 h	20 s	1 363	2	[20]
High School 5	299	34 937	8.99 h	20 s	1 298	2	[20]
Gallery 1	200	5 943	7.80 h	20 s	398	2	[22]
Gallery 2	204	6 709	8.05 h	20 s	393	2	[22]
Gallery 3	186	5 691	7.39 h	20 s	362	2	[22]
Gallery 4	211	7 409	8.01 h	20 s	294	2	[22]
Gallery 5	215	7 634	5.61 h	20 s	967	1	[22]
Cambridge 1	186	3 853 714	6.07 d	20 s	180	1 312	[25]
Cambridge 2	2 536	2 064 114	3.89 d	1 s	5 996	42	[26]
Intel	112	2 448 720	4.15 d	20 s	107	1 326	[25]
Copenhagen Bluetooth	671	458 920	28.0 d	20 s	13 363	2	[29]
Kenya	52	2 070	2.54 d	1 h	43	26	[24]
Diary	49	2 143	4.28 yr	1 d	345	4	[31]
Prostitution	16 730	50 632	6.00 yr	1 d	39 044	1	[32]
St Andrews	25	408 996	74 d	1 s	139	379	[27]
WiFi	18 719	9 094 619	83.7 d	5 m	884 800	6	[41]
Facebook	45 813	855 542	4.28 yr	1 s	183 412	1	[33]
Messages	35 624	489 653	8.27 yr	1 s	94 768	2	[36]
Forum	7 084	1 429 573	8.61 yr	1 s	70 942	2	[36]
Dating	29 341	529 890	1.15 yr	1 s	74 561	2	[35]
Copenhagen Calls	483	10 545	28.0 d	1 s	271	6	[29]
Copenhagen SMS	533	30 380	21.6 d	1 s	320	12	[29]
E-mail 1	57 194	444 160	112 d	1 s	92 442	1	[37]
E-mail 2	3 188	309 125	81 d	1 s	16 220	3	[38]
E-mail 3	986	332 334	1.52 yr	1 s	9 474	3	[39]
E-mail 4	167	82 927	271 d	1 s	1 830	4	[40]
College	1 899	59 835	193 d	1 s	8 608	2	[34]

probabilities, all nodes in the component of the infection seed will be infected. In such a case it would make sense to place a sentinel in the largest component (where the disease most likely starts).

4. Harmonic closeness

Closeness centrality builds on the assumption that a node that has, on average, short distances to other nodes is central [44]. Here, the distance $d(i, j)$ between nodes i and j is the number of links in the shortest paths between the nodes. The classical measure of closeness centrality of a node i is the reciprocal average distance between i and all other nodes. In a fragmented network, for all nodes, there will be some other node that it does not have a path to, meaning that the closeness centrality is ill defined. (Assigning the distance infinity to disconnected pairs would give the closeness centrality zero for all nodes.) A remedy for this is, instead of measuring

the reciprocal average of distances, measuring the average reciprocal distance [45],

$$C_C(i, G) = \frac{1}{N} \sum_{j \neq i} \frac{1}{d(i, j)}, \qquad (2)$$

where $d^{-1}(i, j) = 0$ if i and j are disconnected. We call this the *harmonic closeness* by analogy to the harmonic mean.

5. Harmonic vitality

Vitality measures are a class of network descriptor that capture the impact of deleting a node on the structure of the entire network [46,47]. Specifically, we measure the harmonic closeness vitality, or *harmonic vitality*, for short. This is the change of the sum of reciprocal distances of the graph (thus, by analogy to the harmonic closeness, well defined even for

disconnected graphs):

$$C_V(i, G) = \frac{\sum_{j \in G} C_C(j, G)}{\sum_{j \in G \setminus \{i\}} C_C(j, G \setminus \{i\})}. \quad (3)$$

Here the denominator concerns the graph G with the node i deleted. If deleting i breaks many shortest paths, then $C_C(i)$ decreases, and thus $C_V(i)$ increases. A node whose removal disrupts many shortest paths would thus score high in harmonic vitality.

6. Coreness

Our sixth structural descriptor is *coreness*. This measure comes out of a procedure called k-core decomposition. First, remove all nodes with degree $k = 1$. If this would create new nodes with degree one, delete them too. Repeat this until there are no nodes of degree 1. Then, repeat the above steps for larger k values. The coreness of a node is the last level when it is present in the network during this process [48].

F. Temporal network descriptors
1. Degree

Like for the static networks, in the temporal networks we measure the degree of the nodes. To be precise, we define the degree as the number of distinct other nodes a node in contact with within the data set.

2. Strength

Strength is the total number of contacts a node has participated in throughout the data set. Unlike degree, it takes the number of encounters into account.

3. Up- and downstream component sizes

Temporal networks, in general, tend to be more disconnected than static networks. For node i to be connected to j in a temporal networks there has to be a time-respecting path from i to j, i.e., a sequence of contacts increasing in time that (if time is projected out) is a path from i to j [7,8]. Thus two interesting quantities—corresponding to the component sizes of static networks—are the fraction of nodes reachable from a node by time-respecting paths forward (downstream component size) and backward in time (upstream component size) [49].

4. Temporal statistics

If a node only exists in the very early stage of the data, the sentinel will likely not be active by the time the outbreak happens. If a node is active only at the end of the data set, it would also be too late to discover an outbreak early. For these reasons, we measure statistics of the times of the contacts of a node. We measure the *average time* of all contacts a node participates in; the *first time* of a contact (i.e., when the node enters the data set); and the *duration* of the presence of a
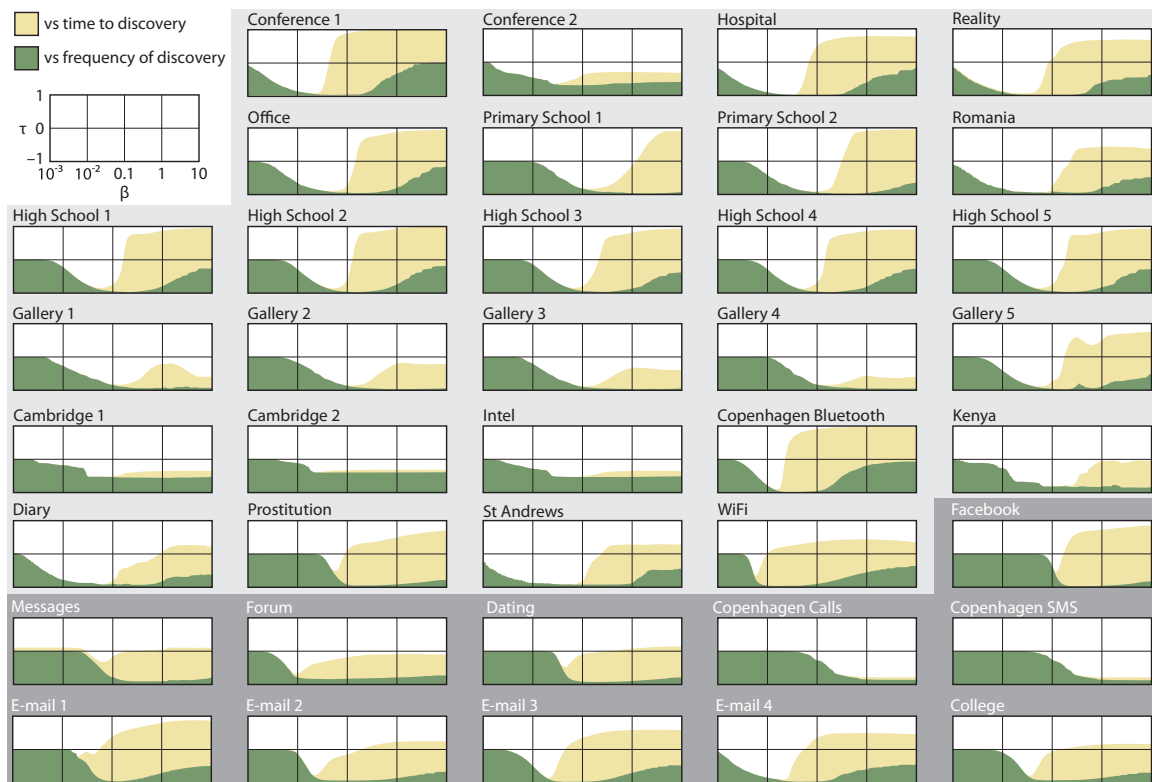

FIG. 2. Kendall τ correlation between the different objective measures for the static networks. For every data set, we show the correlations between time to detection or extinction and the other two objective measures. Since τ for correlations with the frequency of detection is never larger than the correlations with the time to detection, we can highlight the curves by coloring the area underneath (without any point being covered). The β axis is logarithmic.
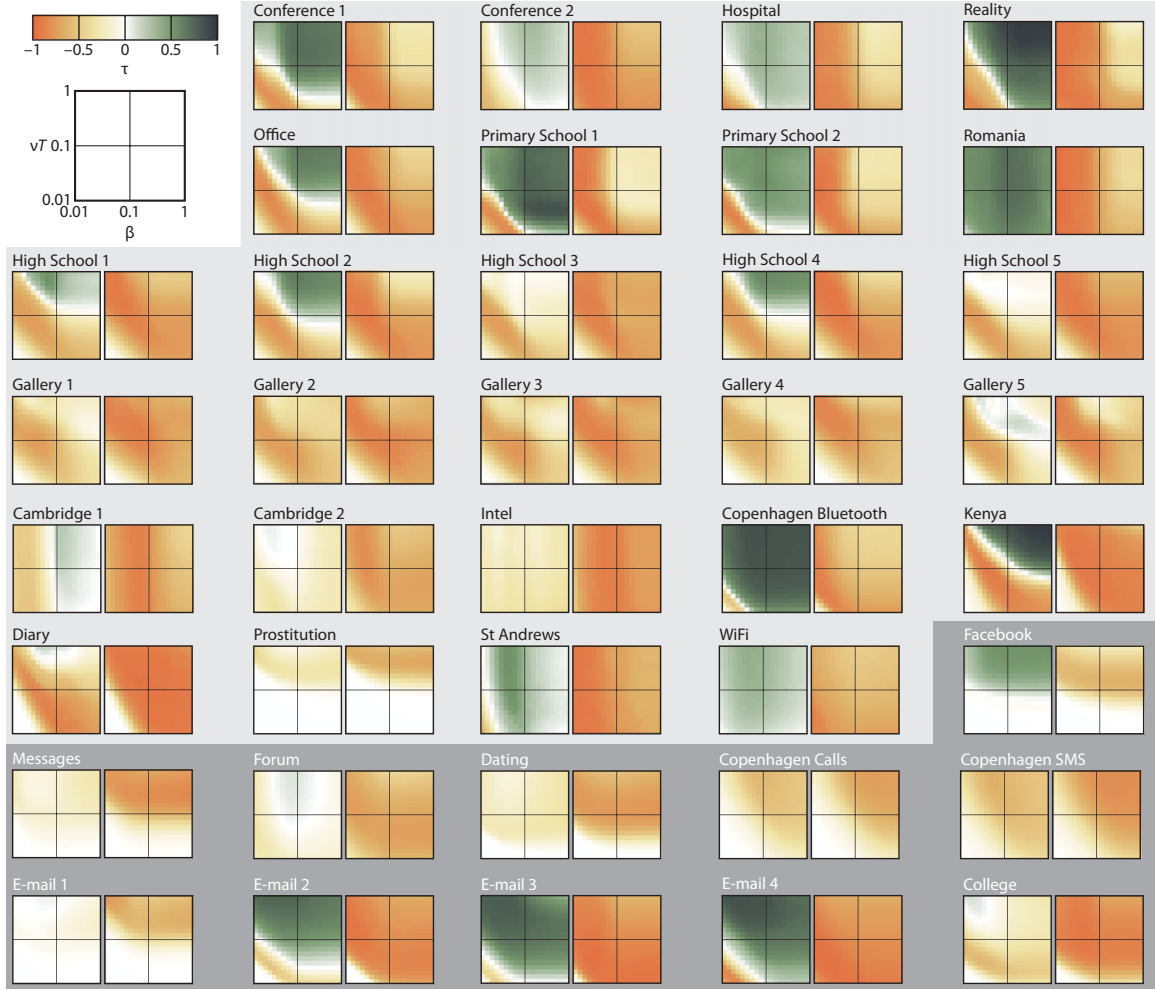
FIG. 3. Correlation between the different objective measures for the temporal networks. For every data set, the left panel shows the Kendall τ values for correlations between time to detection or extinction and time to detection, whereas the right panel shows the corresponding values for time to detection or extinction and frequency of detection. The axes are logarithmic.

node in the data (the time between the first and last contact it participates in).

G. Modified Kendall's τ coefficient

We use a version of the Kendall τ coefficient [50] to elucidate both the correlations between the three objective measures, and between the objective measures and network structural descriptors. In its basic form, the Kendall τ measures the difference between the number of concordant (with a positive slope between them) and discordant pairs relative to all pairs. There are a few different versions that handle ties in different ways. We count a pair of points whose error bars overlap as a tie and calculate

$$\tau = \frac{n_c - n_d}{n_c + n_d + n_t},\qquad(4)$$

where n_c is the number of concordant pairs, n_d is the number of discordant pairs, and n_t is the number of ties.

III. RESULTS

A. Correlation between the objective measures

We start investigating the correlation between the three objective measures throughout the parameter space of the SIR model for all our data sets.

1. Static networks

We use the time to detection and extinction as our baseline and compare the other two objective measures with that. In Fig. 2, we plot the τ coefficient between t_x and t_d and between t_x and f_d. We find that for low enough values of β, the τ for all objective measures coincide. For very low β the disease just dies out immediately, so the measures are trivially equal: all nodes would be as good sentinels in all three aspects. For slightly larger β—for most data sets $0.01 < \beta < 0.1$—both $\tau(t_x, t_d)$ and $\tau(t_x, f_d)$ are negative. This is a region where outbreaks typically die out early. For a node to have low t_x, it needs to be where outbreaks are likely to survive, at least for a while. This translates to a large f_d, while for t_d, it would be beneficial to be as central as possible.

If there are no extinction events at all, t_x and t_d are the same. For this reason, it is no surprise that, for most of the data sets, $\tau(t_x, t_d)$ becomes strongly positively correlated for large β values. The $\tau(t_x, f_d)$ correlation is negative (of a similar magnitude), meaning that for most data sets the different methods would rank the possible sentinels in the same order. For some of the data sets, however, the correlation never becomes positive even for large β values (like *Copenhagen Calls* and *Copenhagen SMS*). These networks are the most fragmented onesm meaning that one sentinel unlikely would detect the outbreak (since it probably happens in another component). This makes t_x rank the important nodes in a way similar to f_d, but since diseases that do reach a sentinel do it faster in a small component than a large one, t_x and t_d become anticorrelated.

2. Temporal networks

In Fig. 3, we perform the same analysis as in the previous section but for static networks. The picture is to some extent similar, but also much richer. Just as for the case of static networks, $\tau(t_x, f_d)$ is always nonpositive, meaning the time to detection or extinction ranks the nodes in a way positively correlated with the frequency of detection. Furthermore, like the static networks, $\tau(t_x, t_d)$ can be both positively and negatively correlated. This means that there are regions where t_d ranks the nodes in the opposite way than the t_x. These regions of negative $\tau(t_x, t_d)$ occur for low β and ν. For some data sets—for example the *Gallery* data sets, *Dating*, *Copenhagen calls*, and *Copenhagen SMS*—the correlations are negative throughout the parameter space.

Among the data sets with a qualitative difference between the static and temporal representations, we find *Prostitution* and *E-mail 1* both have strongly positive values of $\tau(t_x, t_d)$ for large β values in the static networks but moderately negative values for temporal networks.

B. Correlation between objective measures and structural descriptors

In this section, we take a look at how network structures affect our objective measures.

1. Static networks

In Fig. 4, we show the correlation between our three objective measures and the structural descriptors as a function of β for the *Office* data set. Panel (a) shows the results for the time to detection or extinction. There is a negative correlation between this measure and traditional centrality measures like degree or subgraph centrality. This is because t_x is a quantity one wants to minimize to find the optimal sentinel, whereas for all the structural descriptors a large value means that a node is a candidate sentinel node. We see that degree and subgraph centrality are the two quantities that best predict the optimal sentinel location, while coreness is also close (at around -0.65). This in line with research showing that certain biological problems are better determined by degree than more elaborate centrality measures [51]. Over all, the τ curves are rather flat. This is partly explained by τ being a rank correlation
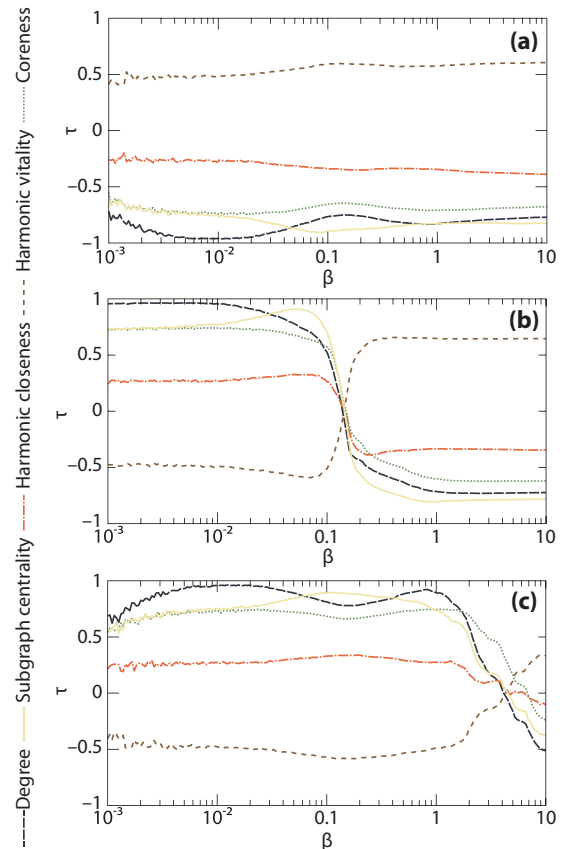

FIG. 4. Correlations between the three objective measures and various quantities describing the static network structure for the *Office* data set. Panel (a) shows results for the time to detection or extinction, (b) shows results for the time to detection, (c) shows results for the frequency of detection.

coefficient:if the rankings do not change (even if the objective measures do), then neither do the τ values.

For t_d [Fig. 4(b)], most curves change behavior around $\beta = 0.2$. This is the region when larger outbreaks could happen, so one can understand there is a transition to a situation similar to t_x [Fig. 4(a)]. f_d [Fig. 4(c)] shows a behavior similar to t_d in that the curves start changing order, and what was a correlation at low β becomes an anticorrelation at high β. This anticorrelation is a special feature of this particular data set, perhaps due to its pronounced community structure. Nodes of degree 0, 1, and 2 have a strictly increasing values of f_d, but for some of the high degree nodes (that all have f_d close to one) the ordering gets anticorrelated with degree which makes Kendall's τ negative. Since rank-based correlations are more principled for skew-distributed quantities common in networks, we keep them. We currently investigate what creates these unintuitive anticorrelations among the high degree nodes in this data set.

Next, we proceed with an analysis of all data sets. We summarize plots like Fig. 4 by the structural descriptor with the largest magnitude of the correlation $|\tau|$. See Fig. 2. We can see, that there is not one structural quantity that uniquely determines the ranking of nodes, there is not even one that dominates over
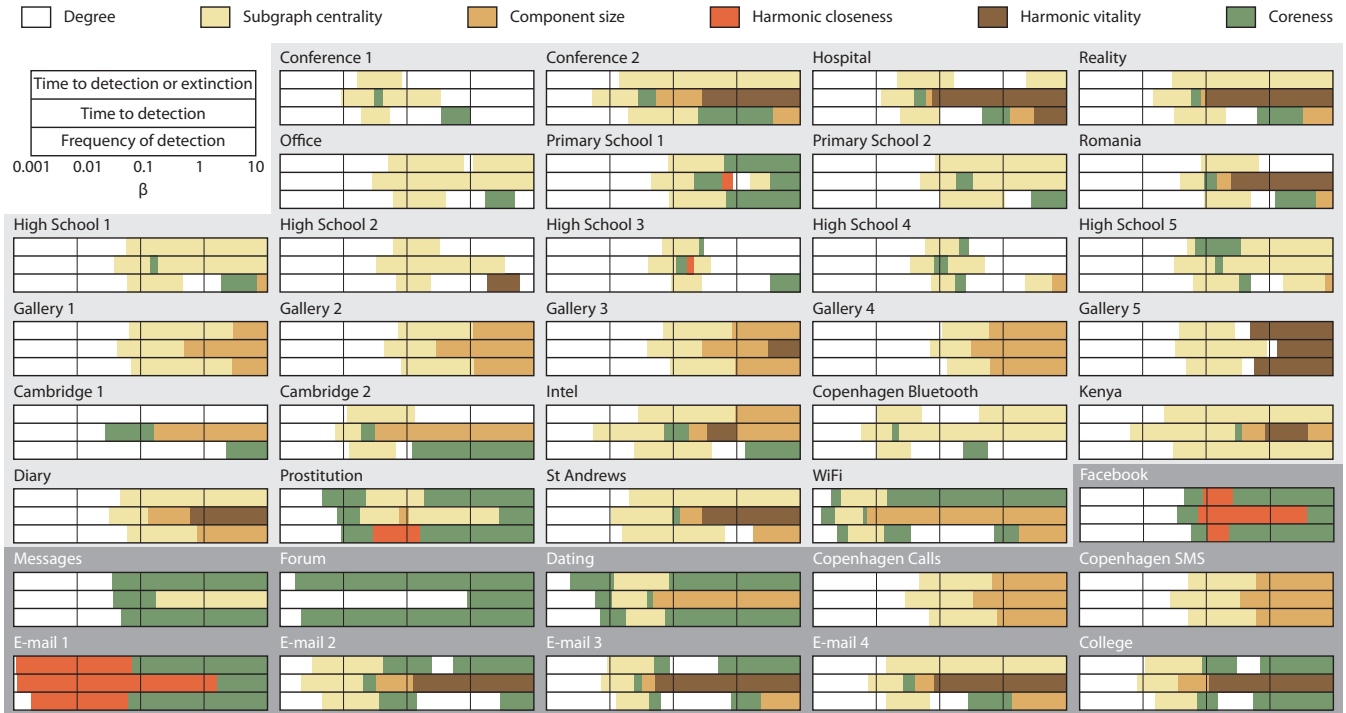
FIG. 5. The strongest correlation between the three objective measures and various measures of the position of nodes for the static networks. Lighter shaded background are data sets of human proximity, the darker background figures indicate data of human communication. The β axis is logarithmic.

the range of β that we investigate. Furthermore, there are some striking patterns:
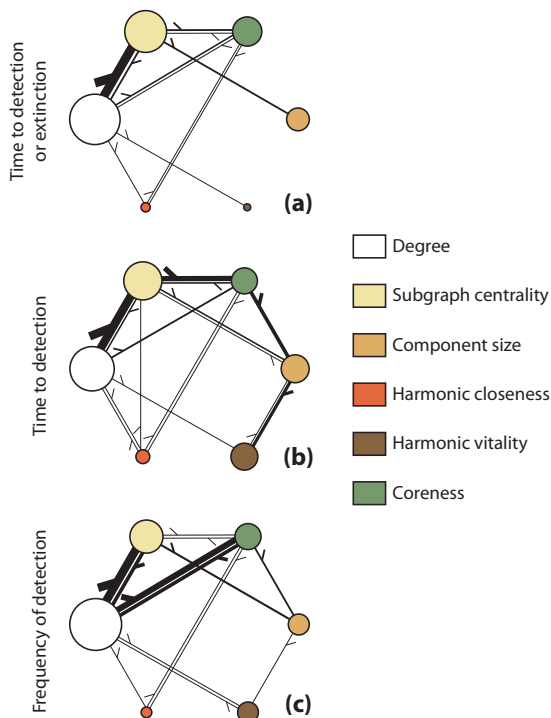

FIG. 6. Transition graphs. The areas of the circles correspond to the frequency of the structural measure in Fig. 5. The widths of the lines are proportional to how many times one measure is succeeded by another as β increases.

(1) Degree is the strongest structural determinant of all objective measures at low β values. This is consistent with Ref. [13].

(2) Component size only occurs for large β. In the limit of large β, f_d is only determined by component size (if we would extend the analysis to even larger β, subgraph centrality would have the strongest correlation for the frequency of detection).

(3) Harmonic vitality is relatively better as a structural descriptor for t_d, less so for t_x and f_d. t_x and f_d capture the ability of detecting an outbreak before it dies, so for these quantities one can imagine more fundamental quantities like degree and the component size are more important.

(4) Subgraph centrality often shows the strongest correlation for intermediate values of β. This is interesting, but difficult to explain since the rationale of subgraph centrality builds on cycle counts and there is no direct process involving cycles in the SIR model.

(5) Harmonic closeness rarely gives the strongest correlation. If it does, it is usually succeeded by coreness and the data set is typically rather large.

(6) Datasets from the same category can give different results. Perhaps *College* and *Facebook* is the most conspicuous example. In general, however, similar data sets give similar results.

The final observation could be extended. We see that, as β increases, one color tends to follow another. This is summarized in Fig. 6, where we show transition graphs of the different structural descriptors such that the size corresponds to their frequency in Fig. 7, and the size of the arrows show how often one structural descriptor is succeeded by another as β is increased. For t_x, the degree and subgraph centrality
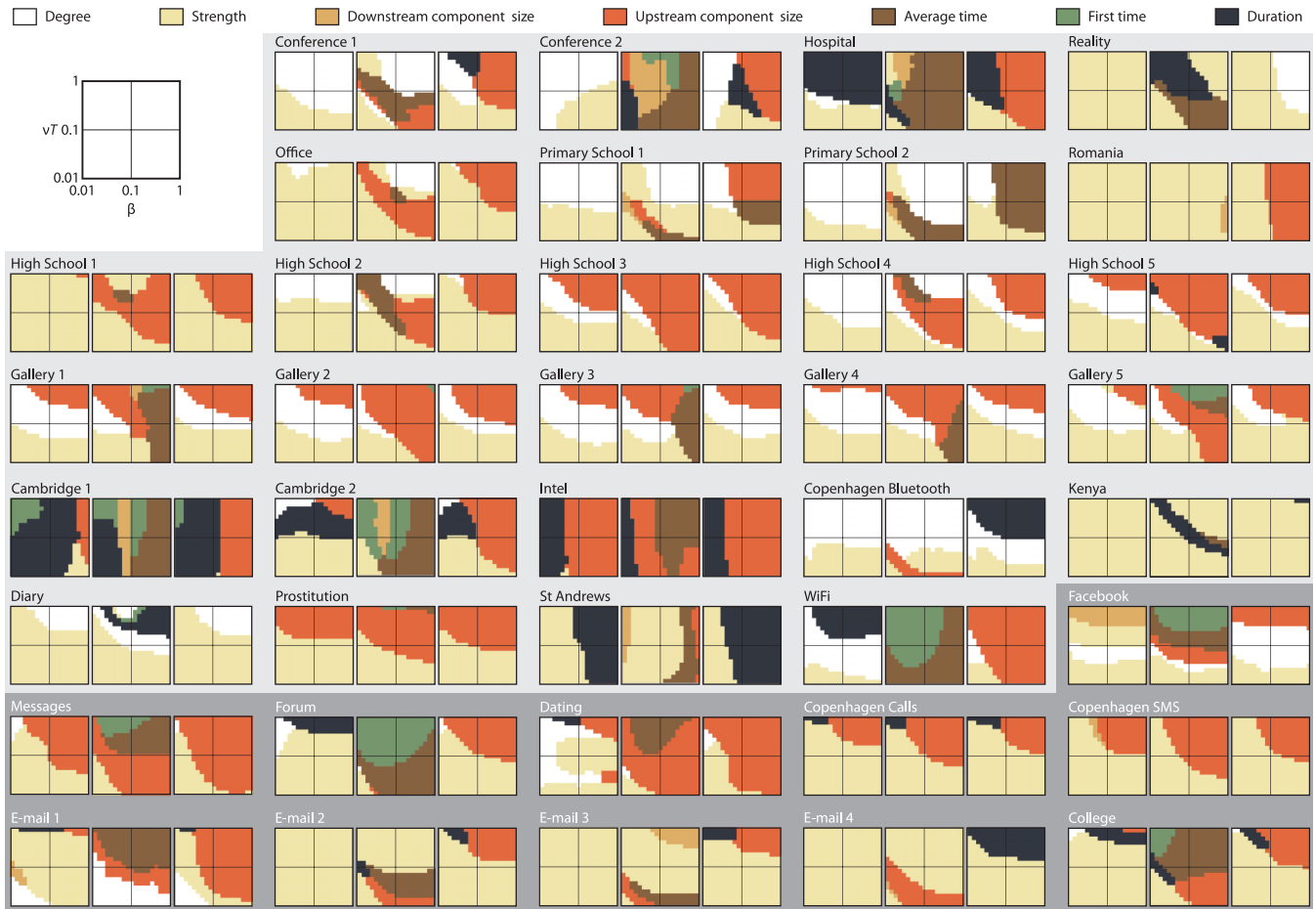
FIG. 7. For every β value, this figure shows the strongest correlation between the three objective measures and various measures of network position for the temporal networks. For each data set, the upper panel shows the results for t_x, the middle panel t_d, and the lower panel f_d. The lighter background shows data sets of human proximity while the darker background is based on human communication data. The axes are logarithmic.

are the most important structural descriptors, and the former is usually succeeded by the latter. For t_d, there is a common peculiar sequence of degree, subgraph centrality, coreness component size, and harmonic vitality that is manifested as the peripheral, clockwise path of Fig. 6(b). Finally, f_d is similar to t_x except that there is a rather common transition from degree to coreness, and harmonic vitality is, relatively speaking, a more important descriptor.

2. Temporal networks

In Fig. 7, we show the figure for temporal networks corresponding to Fig. 5. Just like the static case, even though every data set and objective measure is unique, we can make some interesting observations.

(1) Strength is most important for small ν and β. This is analogous to degree dominating the static network at small parameter values.

(2) Upstream component size dominates at large ν and β. This is analogous to the component size of static networks. Since temporal networks tend to be more fragmented than static

ones [49], this dominance at large outbreak sizes should be even more pronounced for temporal networks.

(3) Most of the variation happens in the direction of larger ν and β. In this direction, strength is succeeded by degree which is succeeded by upstream component size.

(4) Like the static case, and the analysis of Figs. 5 and 7, t_x and f_d are qualitatively similar compared to t_d.

(5) Temporal quantities, such as the average and first times of a node's contacts, are commonly the strongest predictors of t_d.

(6) When a temporal quantity is the strongest predictor of t_x and f_d it is usually the duration. It is understandable that this has little influence on t_d, since the ability to be infected at all matters for these measures; a long duration is beneficial since it covers many starting times of the outbreak.

(7) Similar to the static case, most categories of data sets give consistent results, but some differ greatly (*Facebook* and *College* is yet again a good example).

The bigger picture these observations paint is that, for our problem, the temporal and static networks behave rather similarly, meaning that the structures in time do not matter so much for our objective measures. At the same time, there is

not only one dominant measure for all the data sets. Rather are there several structural descriptors that correlate most strongly with the objective measures depending on ν and β.

IV. SUMMARY AND CONCLUSIONS

In this paper, we have investigated three different objective measures for optimizing sentinel surveillance: the time to detection or extinction, the time to detection (given that the detection happens), and the frequency of detection. Each of these measures corresponds to a public health scenario: the time to detection or extinction is most interesting to minimize if one wants to halt the outbreak as quickly as possible, and the frequency of detection is most interesting if one wants to monitor the epidemic status as accurately as possible. The time to detection is interesting if one wants to detect the outbreak early (or else it is not important), which could be the case if manufacturing new vaccine is relatively time consuming. We investigate these cases for 38 temporal network data sets and static networks derived from the temporal networks.

Our most important finding is that, for some regions of parameter space, our three objective measures can rank nodes very differently. This comes from the fact that SIR outbreaks have a large chance of dying out in the very early phase [52], but once they get going they follow a deterministic path. For this reason, it is thus important to be aware of what scenario one is investigating when addressing the sentinel surveillance problem.

Another conclusion is that, for this problem, static and temporal networks behave reasonably similarly (meaning that the temporal effects do not matter so much). Naturally, some of the temporal networks respond differently than the static ones, but compared to, e.g., the outbreak sizes or time to extinction [53–55], differences are small.

Among the structural descriptors of network position, there is no particular one that dominates throughout the parameter space. Rather, local quantities like degree or strength (for the temporal networks) have a higher predictive power at low parameter values (small outbreaks). For larger parameter values, descriptors capturing the number of nodes reachable from a specific node correlate most with the objective measures rankings. Also in this sense, the static network quantities dominate the temporal ones, which is in contrast to previous observations (e.g., Refs. [53–55]).

For the future, we anticipate work on the problem of optimizing sentinel surveillance. An obvious continuation of this work would be to establish the differences between the objective metrics in static network models. To do the same in temporal networks would also be interesting, although more challenging given the large number of imaginable structures. Yet an open problem is how to distribute sentinels if there are more than one. It is known that they should be relatively far away [13], but more precisely where should they be located?

ACKNOWLEDGMENTS

[1] J. Giesecke, *Modern Infectious Disease Epidemiology* (Arnold, London, 1994).

[2] H. W. Hethcote, SIAM Rev. **42**, 599 (2000).

[3] R. M. Anderson and R. M. May, *Infectious Diseases in Humans* (Oxford University Press, Oxford, 1992).

[4] R. Pastor-Satorras, C. Castellano, P. Van Mieghem, and A. Vespignani, Rev. Mod. Phys. **87**, 925 (2015).

[5] N. Masuda and P. Holme, F1000Prime Rep. **5**, 6 (2015).

[6] N. Masuda and P. Holme, in *Temporal Network Epidemiology* (Springer Nature, Singapore, 2017), pp. 1–16.

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

[8] P. Holme, Eur. Phys. J. B **88**, 1 (2015).

[9] P. Bajardi, A. Barrat, L. Savini, and V. Colizza, J. R. Soc. Interface **9**, 2814 (2012).

[10] N. A. Christakis and J. H. Fowler, PLoS One **5**, 1 (2010).

[11] *Principles and Practices of Public Health Surveillance*, edited by S. M. Teutsch and R. E. Churchill (Oxford University Press, Oxford, 2010).

[12] Y. Bai, B. Yang, L. Lin, J. L. Herrera, Z. Du, and P. Holme, Sci. Rep. **7**, 4804 (2017).

[13] P. Holme, Phys. Rev. E **96**, 062305 (2017).

[14] H. Andersson and T. Britton, *Stochastic Epidemic Models and Their Statistical Analysis* (Springer, Berlin, 2000).

[15] M. S. Bartlett, J. R. Statist. Soc. B **11**, 211 (1949).

[16] Pretty quick code for regular (continuous time, Markovian) SIR on networks, github.com/pholme/sir, accessed May 22, 2018.

[17] Sociopatterns, sociopatterns.org, accessed May 22, 2018.

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

[19] 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.*, PLoS One **6**, e23176 (2011).

[20] R. Mastrandrea, J. Fournet, and A. Barrat, PLoS One **10**, 1 (2015).

[21] P. Vanhems, A. Barrat, C. Cattuto, J.-F. Pinton, N. Khanafer, C. Régis, B.-A. Kim, B. Comte, and N. Voirin, PLoS One **8**, e73970 (2013).

[22] W. van den Broeck, M. Quaggiotto, L. Isella, A. Barrat, and C. Cattuto, Leonardo **45**, 285 (2012).

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

[24] M. C. Kiti, M. Tizzoni, T. M. Kinyanjui, D. C. Koech, P. K. Munywoki, M. Meriac, L. Cappa, A. Panisson, A. Barrat, C. Cattuto *et al.*, EPJ Data Sci. **5**, 21 (2016).

[25] A. Chaintreau, P. Hui, J. Crowcroft, C. Diot, R. Gass, and J. Scott, IEEE Trans. Mobile Comput. **6**, 606 (2007).

[26] J. Leguay, A. Lindgren, J. Scott, T. Friedman, and J. Crowcroft, in *Proceedings, ACM SIGCOMM 2006–Workshop on Challenged Networks (CHANTS), Pisa, Italy, 2006* (ACM, New York, 2006).

[27] G. Bigwood, T. Henderson, D. Rehunathan, M. Bateman, and S. Bhatti, CRAWDAD dataset st_andrews/sassy (v. 2011-06-03), downloaded from http://crawdad.org/st_andrews/sassy/20110603/mobile, 2011.

[28] N. Eagle and A. S. Pentland, Pers. Ubiquitous Comput. **10**, 255 (2006).

[29] A. Stopczynski, V. Sekara, P. Sapiezynski, A. Cuttone, J. E. Larsen, and S. Lehmann, PLoS One **9**, e95978 (2014).

[30] M. Radu-Corneliu, D. Ciprian, and X. Fatos, in *Third International Conference on Emerging Intelligent Data and Web Technologies, September, 2012* (IEEE Computer Society, Piscataway NJ, 2012), pp. 133–139.

[31] J. M. Read, K. T. D. Eames, and W. J. Edmunds, J. R. Soc. Interface **5**, 1001 (2008).

[32] L. E. C. Rocha, F. Liljeros, and P. Holme, Proc. Natl. Acad. Sci. USA **107**, 5706 (2010).

[33] B. Viswanath, A. Mislove, M. Cha, and K. P. Gummadi, in *Proceedings of the 2nd ACM Workshop on Online Social Networks, WOSN '09* (ACM, New York, 2009), pp. 37–42.

[34] P. Panzarasa, T. Opsahl, and K. M. Carley, J. Am. Soc. Inf. Sci. Technol. **60**, 911 (2009).

[35] P. Holme, C. R. Edling, and F. Liljeros, Social Networks **26**, 155 (2004).

[36] F. Karimi, V. C. Ramenzoni, and P. Holme, Physica A **414**, 263 (2014).

[37] H. Ebel, L.-I. Mielsch, and S. Bornholdt, Phys. Rev. E **66**, 035103 (2002).

[38] J.-P. Eckmann, E. Moses, and D. Sergi, Proc. Natl. Acad. Sci. USA **101**, 14333 (2004).

[39] A. Paranjape, A. R. Benson, and J. Leskovec, in *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining, WSDM '17* (ACM, New York, 2017), pp. 601–610.

[40] R. Michalski, S. Palus, and P. Kazienko, in Business Information Systems, in *Proceedings of the 14th International Conference, BIS 2011, Poznan, Poland, June 15–17, 2011*, edited by W. Abramowicz, Lecture Notes in Business Information Processing Vol. 87 (Springer, Berlin, 2011), Vol. 87, pp. 197–206.

[41] Y.-Q. Zhang, X. Li, J. Xu, and A. Vasilakos, IEEE Trans. Syst. Man Cybern. **45**, 214 (2015).

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

[43] E. Estrada and J. A. Rodríguez-Velázquez, Phys. Rev. E **71**, 056103 (2005).

[44] G. Sabidussi, Psychometrika **31**, 581 (1966).

[45] P. Holme and G. Ghoshal, Phys. Rev. Lett. **96**, 098701 (2006).

[46] H. W. Corley and H. Chang, Manage. Sci. **64**, 022305 (1974).

[47] D. Koschützki, K. Lehmann, L. Peeters, S. Richter, D. Tenfelde-Podehl, and O. Zlotowski, in *Network Analysis: Methodological Foundations* (Springer, Berlin, 2005), pp. 16–61.

[48] F. Buckley and F. Harary, *Distance in Graphs* (Addison-Wesley, Boston, 1990).

[49] P. Holme, Phys. Rev. E **71**, 046119 (2005).

[50] M. G. Kendall, Biometrika **30**, 81 (1938).

[51] G. Boldhaus, F. Greil, and K. Klemm, Theory Biosci. **132**, 17 (2013).

[52] S. Janson, M. Luczak, and P. Windridge, Random Struct. Algorithms **45**, 726 (2015).

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

[54] K. Hock and N. H. Fefferman, Ecol. Complexity **12**, 34 (2012).

[55] E. Colman, K. Spies, and S. Bansal, BMC Infect. Dis. **18**, 219 (2018).