SCIENTIFIC REPORTS

OPEN

Information content of contact-pattern representations and predictability of epidemic outbreaks

Received: 27 March 2015
Accepted: 27 August 2015
Published: 25 September 2015

Petter Holme

To understand the contact patterns of a population—who is in contact with whom, and when the contacts happen—is crucial for modeling outbreaks of infectious disease. Traditional theoretical epidemiology assumes that any individual can meet any with equal probability. A more modern approach, network epidemiology, assumes people are connected into a static network over which the disease spreads. Newer yet, temporal network epidemiology, includes the time in the contact representations. In this paper, we investigate the effect of these successive inclusions of more information. Using empirical proximity data, we study both outbreak sizes from unknown sources, and from known states of ongoing outbreaks. In the first case, there are large differences going from a fully mixed simulation to a network, and from a network to a temporal network. In the second case, differences are smaller. We interpret these observations in terms of the temporal network structure of the data sets. For example, a fast overturn of nodes and links seem to make the temporal information more important.

Background and problem statement

Epidemics of infectious disease are complex phenomena involving processes at different scales. The smallest and fastest processes happen in the bodies of the infected people, as the pathogen enters and colonize the host. The largest and slowest processes involve the evolution (or rather co-evolution) of the pathogen and hosts and the development of new treatments. Intermediate to these are the movements of infected hosts and the response of society to an emergent outbreak. In this work, we consider the problem of predicting an outbreak based on (partial) knowledge of the small and intermediate scale processes, while treating evolutionary processes as constant. In other words, we assume that we have some understanding of the pathogen (and pathogenesis), the evolving outbreak, and the way people move and come in contact in such a way that the disease can spread. This is the type of challenge modelers face when there is an outbreak of a new pathogen, or a pathogen in a previously unaffected population[1,2].

To be more specific, we assume the small-scale processes are well modeled by a compartmental model—a scheme dividing the population into categories with respect to the disease and assigning transition rules between the classes. We will focus on the Susceptible-Infectious-Recovered (SIR) model—the canonical model of diseases that makes infected persons immune upon recovery[2]. (For details about the simulation, see the *Methods* section.) A compartmental model needs to be complemented by a model of how people meet and interact, i.e. the contact patterns. We will compare three ways, or levels, of including contact information. The first way is to assume that we know, or are able to model, who is in contact with whom, and also the time of the contacts. We refer to this level of capturing the contact structure as a *temporal network* representation[3–5]. The next level is a *static network*[6–8] representation where we assume that we know who that can be in contact with whom, but nothing about when the contacts happen. The

Department of Energy Science, Sungkyunkwan University, Suwon 440-746, Korea. Correspondence and requests for materials should be addressed to P.H. (email: holme@skku.edu)

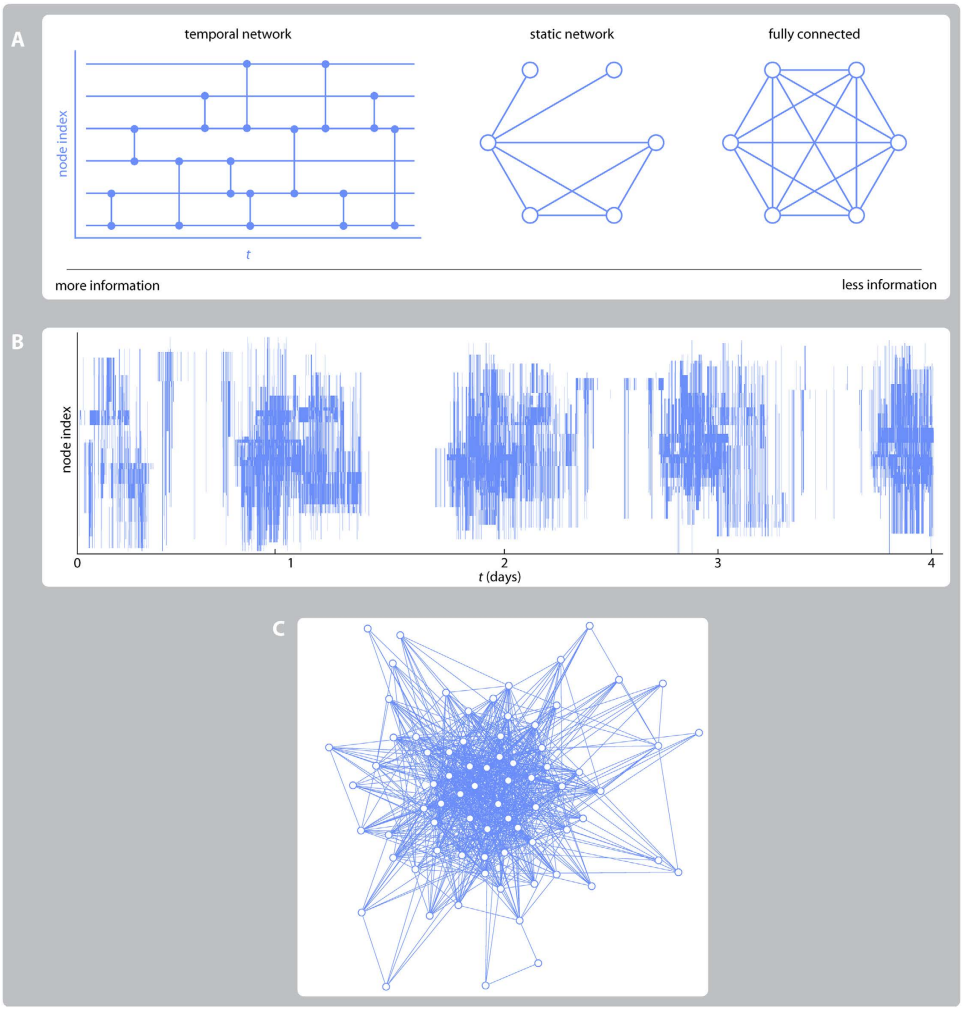

Figure 1. Illustration of the representations of contact patterns containing different levels of information. (**A**) shows the three levels, with respect to information content, of contact representations. The temporal network is visualized using a time line of nodes. If two nodes are in contact at some time, then there is a vertical line at that time. (**B**) gives a real world example of the time-line plot of the temporal network representation of the *Hospital* data. The horizontal lines of (**A**) are, however, omitted. The indices are chosen to minimize the total length of vertical lines. (We chose the *Hospital* data as an example because it has very clear temporal structures, with a conspicuous diurnal pattern.) (**C**) shows the corresponding projected static network of node pairs with more than five contacts.

last level, with the least information content, is a *fully mixed* case[2] where everyone is equally likely to be in contact with anyone else at any time. In terms of networks, this is equivalent to a fully connected network. The three levels of information content are illustrated in Fig. 1. The question of this paper is how including more information about the contacts—going from the fully mixed, to the static, and then to the temporal network representation—changes different aspects of our ability to predict the final outcome of the epidemics. We investigate not only the predicted final outbreak size given there is an outbreak from an unknown node, but also given the current state of an ongoing outbreak at a specific time (the *breaking time*)[9], i.e. assuming more knowledge. The histogram of final outbreak sizes defines what we call unpredictability, or outbreak diversity. There are of course many ways of conceptualizing predictability and defining measures for it, but a broad histogram of outbreak sizes means that there is an inherent stochasticity in the outbreaks that makes it hard to predict the final outcome.

We use empirical contact sequences as the underlying contacts structure for our simulations. We include them either as they are (like temporal networks), or reduce them to static networks, or to fully mixed models. When we project out information, we attempt to keep as much information as possible (cf. ref. 10—so the fully mixed models keep the overall contact frequency of the original data and the same number of contacts will take place over the static networks as in the original data). The seed is chosen randomly and assumed to enter the data at the beginning of the infectious period. Then we scan the entire parameter space of the SIR model on these three representations of the contact patterns. We

	Number of individuals	Number of contacts	Sampling time	Time resolution
Conference	113	20,818	2.50d	20 s
Gallery	200	5,943	7.80 h	20 s
Prostitution	16,730	50,632	6.00y	1d
Hospital	75	32,424	96.5 h	20 s
Reality mining	64	26,260	8.63 h	5 s
School	236	37402	8.61 h	20 s

Table 1. The basic statistics of the data sets.

note that there are different imaginable versions, or scenarios, of this simulation set-up, each probably giving slightly different results. Ultimately our study concerns one particular scenario, described in more detail in the *Methods* section.

Previous studies. The fully mixed case is by far the most studied disease-spreading framework. There are several textbooks and review papers discussing their use. We recommend ref. 2 as an introduction. If one has no information about human contact patterns, we have to treat each agent the same, which leads to the fully mixed approach. But on the other hand, we almost always do have more information—for example, we know that there is a broad distribution of the number of sexual partners, which affects the spread of e.g. HIV[11]. The static network paradigm has been around for at least two decades and profoundly influenced theoretical epidemiology[8]. In addition to making predictions more accurate, one major contribution has been to put an emphasis on the different roles and importance of individuals. Networks provide a framework to explain phenomena such as super spreaders or to find the optimal set of people to vaccinate or quarantine[8]. Temporal network epidemiology[4] is the youngest of these branches of theoretical epidemiology as categorized by their representation of contact patterns. It has mostly been a computational endeavor. This is probably because the many types of structure in temporal data make it hard to study analytically (ref. 12 is a notable exception). There are several studies showing that including time in the contact representations does make a big difference[13–15] in the predicting outbreak sizes. One conclusion is that bursty time series of contacts slow down spreading processes[15]. Another observation is that the birth and death of nodes and links are even more important for disease spreading[16]. A few studies investigate how to exploit the temporal structure to mitigate outbreaks[17,18].

Another line of research recognizes that the contacts are not independent of the disease itself. People would change their contact patterns if they become infected or perhaps just from awareness of the outbreak[19]. In our work, we ignore to model this effect and focus on the impact of the structure of the contact patterns *per se*. There are some other extensions of network theory—spatial[20] and multilayer networks[21]—that are getting increasingly interesting for epidemiology.

The previous work most similar to ours is probably ref. 9 where we look at the decay of unpredictability (defined in a similar way to this paper) as a function of time. In that paper, we investigated (using network models) how static network topology influences this decay.

Empirical data. Our starting point is empirical data sets of temporal human proximity networks—records of two persons being close to each other, and when these contacts happen. Any non-vector-borne infection whose pathogens cannot survive for extended periods outside a host do spread over proximity networks. However, the exact requirement for two persons to be close enough, and the exposure to be long enough, for the disease to spread varies for different diseases. Human proximity data is, however, hard to obtain at a resolution enough to model the epidemics of a specific disease. Instead of focusing on a particular disease (i.e. fine tuning the SIR parameter values to this disease), we scan the entire parameter space and thereby study features general to all SIR type diseases. We list the sizes, sampling durations, etc., of the data sets in Table 1.

One of our data sets comes from the Reality mining study[22] (*Reality*). In this data set, contacts within a cohort of university students were recorded by the Bluetooth devices of their smartphones. The range of such devices is between 10 and 15 meters. To be able to compare our results to other studies, we use a reduced set of contacts from this data set—the same as in refs 16,23.

Another group of proximity data comes from the Sociopatterns project (sociopatterns.org). These data sets are gathered from groups of people wearing radio-frequency identification sensors. Such devices record a contact if two sensors are no further than 1–1.5 m, and the wearers are facing each other. One of these datasets come from the attendees of a conference[24] (*Conference*), another from a school (*School*)[25], another from a hospital (*Hospital*)[26] and yet another from visitors to a gallery (*Gallery*)[27]. The *Gallery* data set comprises 69 days where we use the first three. *School* covers two days and we use both.

A different type of proximity data set that we also study comes from self-reported sexual contacts between female prostitutes and male sex buyers[28]. We call this data set *Prostitution*.

Results

Now we turn to the results of our analyses. For every data set, our raw output is a four dimensional array of values—a histogram of outbreak sizes as a function of the breaking time and the two parameter values of the SIR model. Of course we need to simplify this output by projecting out different dimensions. For further simplification, we will mostly present the results for one data set in the paper and leave the rest for the supplementary information. Which data set that we chose depends on what feature we will highlight in the discussion. In other words, we do not try to show representative results (which is anyway hard to do objectively), but those that help the discussion of the features of our data collection.

There are a few different versions of the SIR model. We assume that a susceptible individual meeting an infectious would make the susceptible infectious with a probability λ. Then the infected node stays infectious for δ timesteps. (See ref. 29 for further motivations of this version.) We sample 20×20 data points growing exponentially from 0.001 to 1 in both δ and λ (δ is normalized by the sampling time of the data). The exponentially series of points enables us to use the same grid for all data sets. If we tailored the grid for the individual data sets we could get a higher resolution, but for the purposes of this paper our method suffices. One can think of different ways to include the empirical contacts into the SIR simulation. Our approach is grounded on two assumptions. First, we assume that the disease spreads only within the recorded set of contacts. An alternative to this would be use the datasets to create a model for the contact patterns, and then to use this model to generate the contacts for the disease simulation. This would be challenging (since we have rather few datasets, all with rather distinctive structures, it is hard to say what the general features are) but an interesting direction for the future. Second, we assume every node has the same probability of introducing the disease to the population. Since we do know more about the nodes than the contacts they are in, and we do not have any method to translate this information to the probability of being the source of the outbreak, we have to treat all nodes as equal.

Predicting outbreak sizes with no knowledge about who is infected. In this section, we investigate how the three levels of representations affect the predicted outbreak sizes given no knowledge about who is infected. This type of comparative study has been done previously to show the effects of including (static) network information[1,7,30] and temporal information[14–18]. However, to our knowledge, this is the first time all three levels of representations are considered simultaneously.

In Fig. 2, we plot the average fraction of nodes that are infected during the outbreak for the *Gallery* data set. Both the static network representation and the fully connected picture are rather different from the results for the temporal networks. Briefly stated, without the temporal component, the simulations overestimate the outbreak sizes (this was also observed in ref. 31). One factor is of course the reduced reachability[3] in the temporal networks—the fact that you cannot reach every individual from every other, even though a path in the static network of aggregated contacts could connect them. This effect is more than a question of the existence of such time-respecting paths—assume the existence of such path from i to an important spreader j hinges on one contact, then chances are high that the outbreak will not reach j from i, and if the important node j is not infected this might reduce the average outbreak sizes much.

The difference between the fully connected and static network is a little bit smaller. Still, among the data sets we test, *Gallery*'s results are most affected by the static network structure. Other datasets, such as *Hospital*, *Conference* and *School*, are more densely connected and thus the spreading is faster and, in effect, more similar to the fully mixed case (cf. Fig. 1C). The static *Gallery* network is stretched out as an effect of time—visitors come, spend some time at the gallery and leave so early visitors would not meet late visitors. In the large-N limit, the static network structure can make a huge difference (the vanishing epidemic threshold for scale-free networks to mention one example[8]). With an exception for the *Prostitution* data, we draw similar conclusions from the other data sets (see Supplementary Fig. S1)—first, the temporal structure makes a larger difference than that the static network structure; second, including this structure makes the outbreaks smaller.

Outbreak diversity and the approach to high predictability with knowledge about who is infected. The average value of the outbreak sizes is of course only one type of result that epidemic simulations can give. They can also predict dynamic quantities such as the early incidence (number of new cases per time)[32] or the extinction time[33,34], and also distributions of outbreak sizes and times. In this work, we will look further at the distribution of predicted outbreak sizes assuming that we know the state of the outbreak—who is susceptible, infectious or recovered and when the infectious people were infected—at any time t. Given this information, the SIR model gives a distribution of outbreak sizes. Once an individual is infected, its final contribution to the outbreak size is determined. Thus, as t increases, the distribution will gradually gather all its weight at zero. How this approach to having total predictability unfolds can vary much. In Fig. 3, we illustrate our method to investigate this. In parallel to a *master run* of the outbreak simulation (the thick line in the plots), we use the state of the system as the seed for 10^3 independent *auxiliary runs*. From an auxiliary run i starting at the breaking time, we measure the fraction of individuals eventually infected Ω_i. Then we measure histograms of $\Delta\Omega(t) = |\Omega_i - \Omega_j|$ over all pairs of auxiliary runs $i \neq j$ starting at time t, and all 10^4 master sequences. $\Delta\Omega$ measures predictability in the following sense: if $\Delta\Omega(t)$ is narrowly distributed around zero, then one can use the observations from previous outbreaks of the same disease (or another disease of the same λ and δ) to accurately predict the final outbreak size.

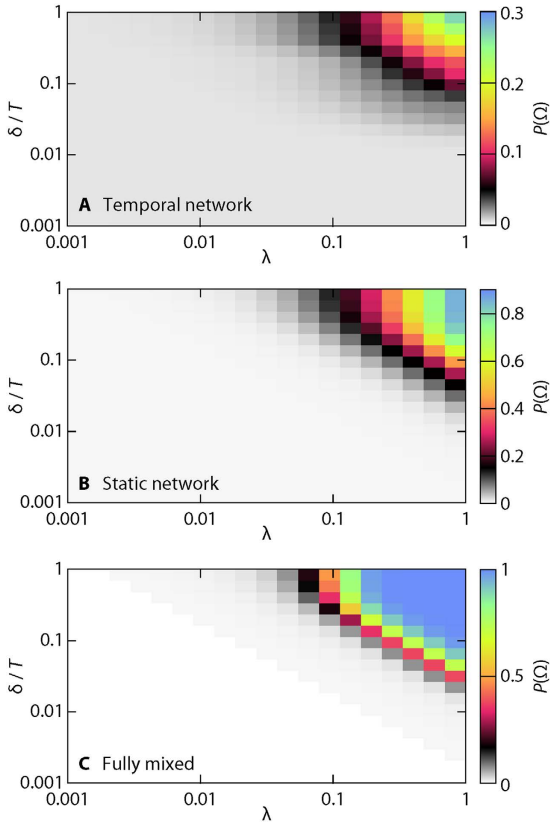

Figure 2. Outbreak sizes for the three representations of contact patterns for the *Gallery* data. (**A**) shows the fraction of recovered individuals at the end of the outbreak for the temporal network data. (**B**) shows the corresponding plot for the static network of aggregated contacts. (**C**) is the plot for the fully connected network. The corresponding plots for the other data sets can be found in Supplementary Fig. S1.

In Fig. 4, we show a number of examples of $\Delta\Omega$ histograms as functions of time (all from the *Hospital* data set). For the temporal network data, there is a fundamental difference between large δ and small λ on one hand, and large λ and small δ on the other. In the former case (Fig. 4A being a typical example), the deviations are continuously decaying (with the peak always around zero). The decay accelerates with t. As δ decreases, or λ increases, the histogram loses its unimodal shape. Typically it turns into a bimodal distribution as seen in Fig. 4B. If δ decreases, or λ increases, further, then the bimodal distribution will split into more, and more well defined, peaks Fig. 4C. This situation, however, exists only in a small fraction of the parameter space. In Fig. 4C, $\lambda = 0.23$ and $\delta = 0.11$. For the next δ-value we measure ($\delta = 0.16$), several of the largest peaks are gone (while some others remain almost identical).

Our interpretation of the above observations is that, if the transmission probability is very small, there is a fairly constant chance for the outbreak to die out. If it was exactly constant, it would give an exponential distribution of extinction times (and probably of $\Delta\Omega$ too). This has been observed before for the SIR model on static networks[32,33] and on simplified models of temporal networks[12] (it is indeed true for our simulations too—see Fig. 4E). The situation for higher transmission probability can be described as a transient when the outbreak can either die or spread. Once it takes off, it behaves rather deterministically[34] (at least in the limit of large population size). This situation can be seen in Fig. 4E, representative of the static networks and fully mixed simulations (that, as we argue below, are similar because the static network is sufficiently dense). This process results in a bimodal distribution. Increasing the transmission probability further while lowering the disease duration makes the process yet more deterministic. At the same time, it also reduces the number of possible outbreak trees. The question, in this parameter region, is not whether there will be an outbreak or not, but which one of a few possible outbreak scenarios that will happen. These few possibilities shows as peaks (or rather lines) in Fig. 4C,D.

Interlude: temporal network structure of the data sets. The main theme of this article is to understand the effects of the level of information content of the contact representation on the deviations from the predicted outbreak sizes. For the discussion in the next section, however, we will need a bit more nuanced picture of the temporal network structure of the data sets. Here we examine three different classes of measures of temporal network structure—those characterizing the static network,

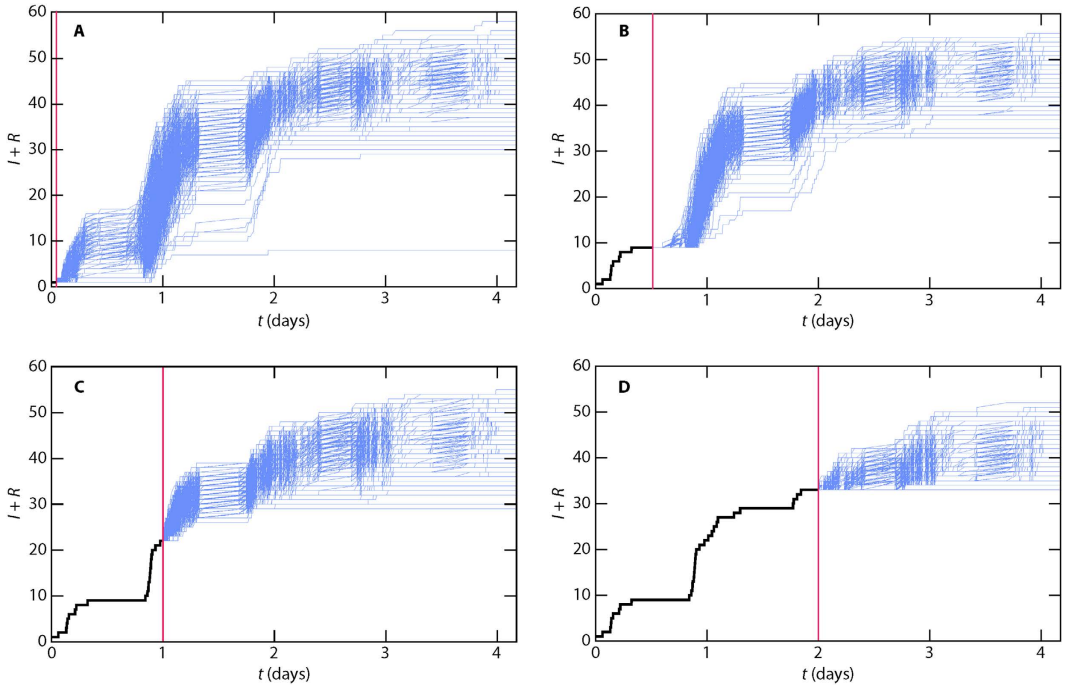

Figure 3. Example of continuations of outbreak trajectories given the state of an outbreak at certain breaking times *t*. This figure illustrates our method to measure the deviation from a predicted outbreak size. The thin lines show 1000 possible future trajectories from the breaking point (indicated by the horizontal line). The thick line shows the trajectory actually taken up to the breaking time. The simulations are from the temporal network representation of the *Hospital* data with parameter values $\delta = 0.6$ and $\lambda = 0.1$.

those characterizing the time series of contacts of individuals and pairs of individuals, and finally those characterizing long-term trends in the activity in the data set.

To summarize the static network structure, the first quantity we study is the coefficient of variation of the degree distribution c_k

$$c_k = \sigma_k / \bar{k} \tag{1}$$

where σ_k and $\bar{k}$ are the standard deviation and mean of the degree—the total number of others that i is in contact with during the sampling. It is known that a heterogeneous degree distribution makes the spreading faster and further reaching[8,35]. The coefficient of variation is a dimensionless measure of the heterogeneity of a distribution.

Another static network measure that is known to affect disease spreading is the clustering coefficient C.

$$C = \frac{3n_{\text{triangle}}}{n_{\text{triplet}}} \tag{2}$$

where n_{triangle} is the number of triangles in the graph of aggregated contacts, and n_{triplet} is the number of triplets (a subgraph of three connected vertices, not necessarily a triangle)[35]. In general, a high value of this coefficient slows down the spreading[36]. This is quite intuitive. Imagine a triangle, where one node infects the two others. Now the link between the secondary infected nodes is superfluous for the disease spreading and it would have benefitted the spreading if was connected to some distant node instead.

The first quantities investigating the temporal aspects are the average time of the presence of nodes d_N and links d_L. To be specific, we define the time of presence as the time between the first and last contact. If one take a longer perspective than the sampling time, it will be an approximation, since the last contact does not necessarily mean that a node or link became inactive. However, ref. 15 indicates that for these data sets the above approximation is not so grave (ref. 15 studies five datasets in common to this paper).

There has been a good deal of interest in how the distribution of times between contacts affects spreading processes. If this is the only temporal structure present in the data, it is known to slow down epidemic spreading[15]. However, ref. 16 argues that birth and death of nodes and (more closely related to d_N and d_L) are more important for disease-type spreading in empirical data sets.

The final two structural measures try to capture a property that sets *Prostitution* aside from the other data sets, namely that the overall activity is increasing through the sampling period. We measure the

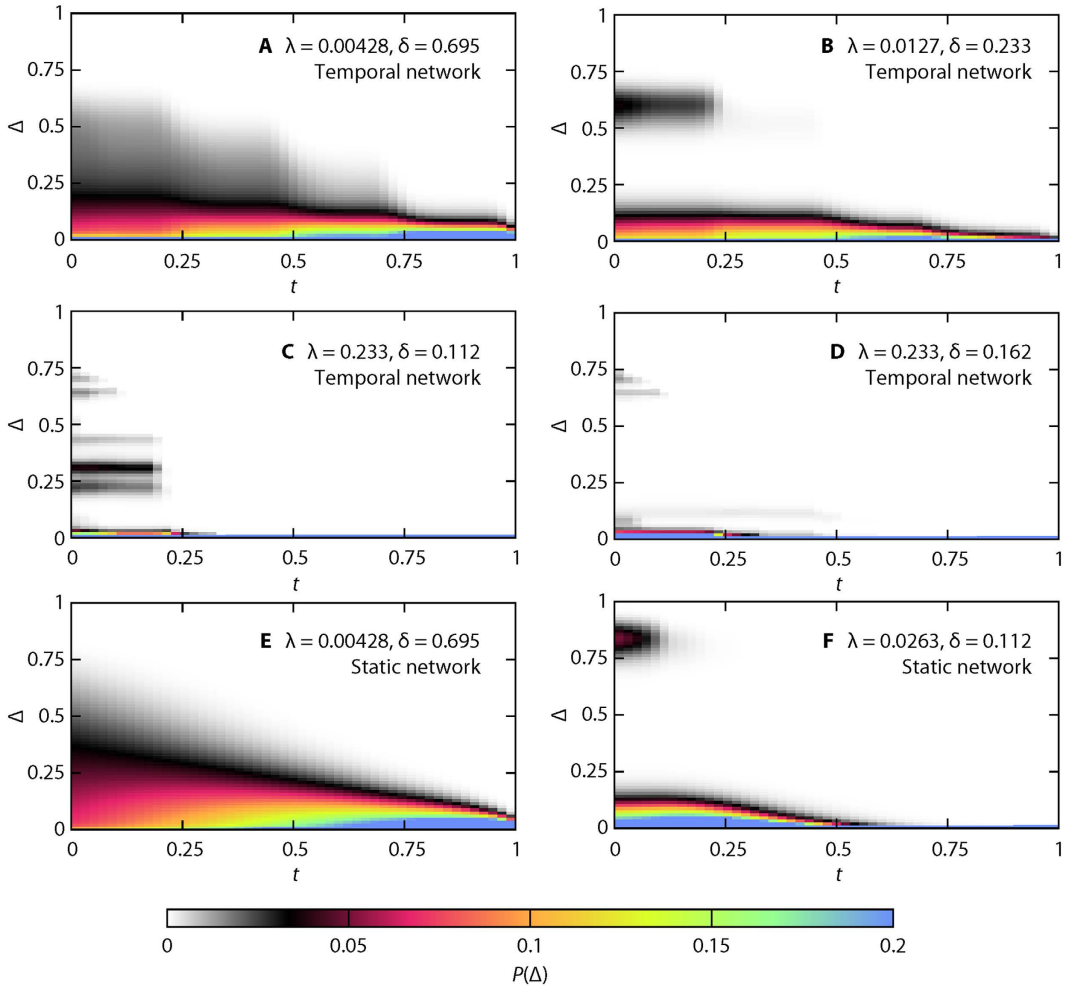

Figure 4. Time evolution of deviation from other outbreaks (*Hospital* data). Panels (**A–D**) shows data from the temporal networks while E and F are for the static networks. A shows a typical plot for low-λ and large-δ values. B shows a bimodal histogram for intermediate λ. (**D,E**) represent large-λ and low-δ. Although the change in δ is not that large between D and E, the pattern of the deviations is. E shows the a plot for the static network representation with the same parameter values as panel (**A**). (**F**) shows a typical bimodal configuration for the static network case (corresponding to panel (**B**), but for slightly different parameter values).

fraction of nodes f_N and links f_L that are present in the data set at half the sampling time[37]. In data sets that sample a growing population, one would expect these quantities to be rather low. If the links are stable and the contacts frequent (the time between them short compared to the sampling time), then f_N and f_L are be large.

The results for the above analysis are summarized as radar plots in Fig. 5. We see that *Prostitution* is indeed very different from the others—it has a more heterogeneous degree distribution, it has (as expected) much lower f_N and f_L, and it has $C = 0$ (since it is a bipartite network). Among the other networks, *Gallery* is the most special as it has very low d_N and d_L values—not surprising, since it samples gallery visitors coming and going during the sampling period.

Time evolution of the predicted outbreak diversity. Next, we look at statistics summarizing histograms like Fig. 4 for all parameter values. We measure the average (Fig. 6) and maximum (Fig. 7) values of the deviation $\Delta\Omega$ of the histograms of the predicted Ω given the state at t. One can think of other summary statistics, but as we will see, we can draw some conclusions that generalize over both the average and maximum. The first observation from Figs 6 and 7 is that there is more complex structure in the curves of the temporal networks. This is no surprise since the static network and fully mixed cases have a time-invariant overall activity. A second observation is that the decay of the unpredictability (a.k.a. outbreak diversity) $\Delta\Omega$ is not extremely fast for any of the data sets and summary statistics. At $t = 0.2\ T$, i.e. at 20% of the sampling time, $\Delta\Omega$ has rarely decreased to less than 20% of its original value. This should be seen in the context of compartmental models on networks being highly predictable in the

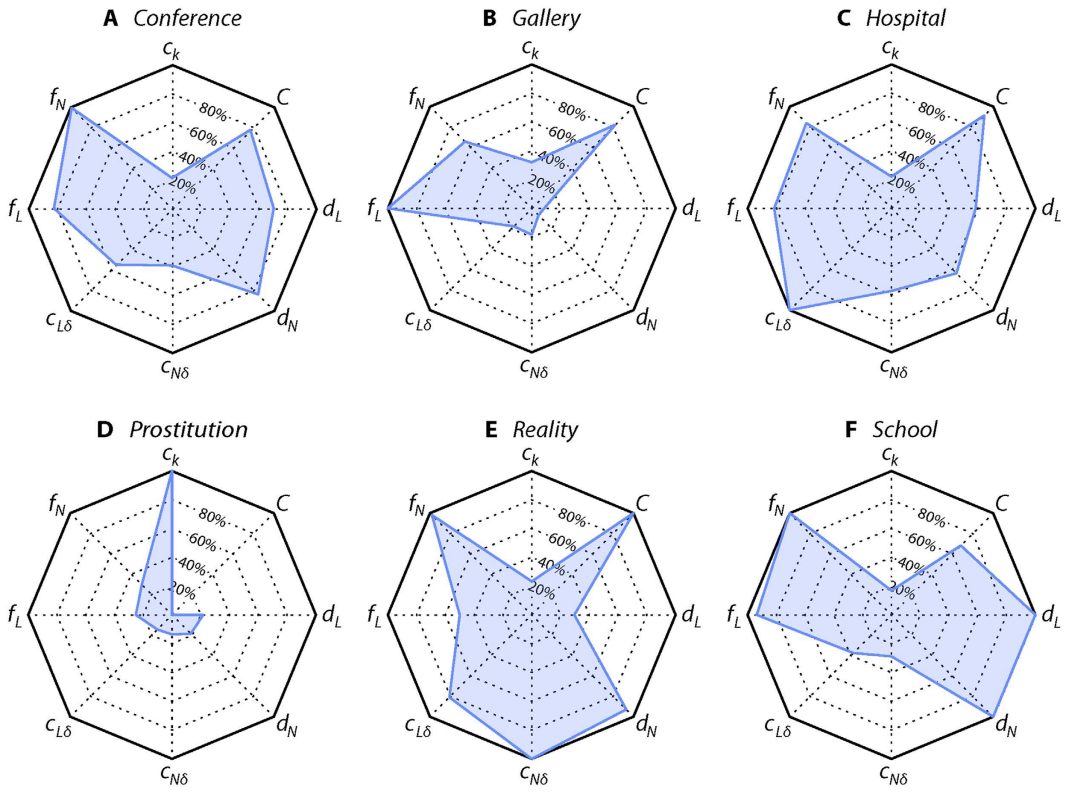

Figure 5. Radar plots summarizing the temporal network structures. The radial component of the areas gives the relative value of the quantity compared to the maximum in among the six data sets. The eight quantities are explained in detail in the *Methods* section. They (and their maximal values) are as follows: c_k—coefficient of variation of the degree distribution (maximum 2.24 for *Prostitution*); C—clustering coefficient (maximum 0.644 for *Reality*); d_L—average duration of links (maximum 0.404T for *School*); d_N—average duration of nodes (maximum 0.938T for *School*); $c_{N\delta}$—node burstiness (maximum 11.1 for *Reality*); $c_{L\delta}$—link burstiness (maximum 15.8 for *Hospital*); f_L—fraction of links present at $T/2$ (maximum 0.783 for *Gallery*); f_N—fraction of nodes present at $T/2$ (maximum 0.987 for *Conference*).

sense that outbreaks either die early or converge to deterministic quantities[38]. Our added insight is that even though the latter observation is true, the convergence may be slow. The *Prostitution* data (Figs 6D and 7D) is a bit different since the values of $\Delta\Omega$ are very low. Probably, the relatively short node and link durations, and the time evolution of the data (reflected in the low f-values) accentuate high predictability for the temporal networks further. Furthermore, we see that at $t = 0$ (i.e. with only the seed node known), temporal networks are usually least unpredictable. *Prostitution* is a big exception for the average $\Delta\Omega$ (Fig. 6D) and *Reality* for the maximal Ω (Fig. 7E).

Yet an observation is that the fully mixed case often starts with a higher average (or maximal) $\Delta\Omega$ compared to the static network case, but then decays faster so that for larger t the fully mixed case has smaller outbreak diversity. This tendency is strongest for the networks with most heterogeneous degree distributions (*Prostitution* and *Gallery*). Other than that, it is hard to speculate about the mechanisms for this observation without using a model to tune the network structure (which is an interesting future project, beyond the scope of the present paper).

Our final, and perhaps most interesting observation, is that there is no clear relation between the temporal network representation on one hand and the other two representations on the other hand. For *Gallery* the temporal network representation is more predictable (have smaller outbreak diversity), for *Reality* it is less predictable. We think that small d_N and d_L (like for *Prostitution* and *Gallery*) could, in general, implicate that adding temporal information increases the predictability (as observed above) much. The reason is that then the order of the appearance of nodes and links will matter more. The contacts will then work more like a river system where water flows from higher elevation to lower (or, in our case, from earlier nodes and links to later). Finally, we note that for some data sets (*Conference* and *Prostitution*) the ranking of the representation changes over time. In general, the difference by adding information about time (i.e. going from a static to a temporal network representations) is smaller for $\Delta\Omega$ than Ω (Fig. 2 and Supplementary Fig. S1).

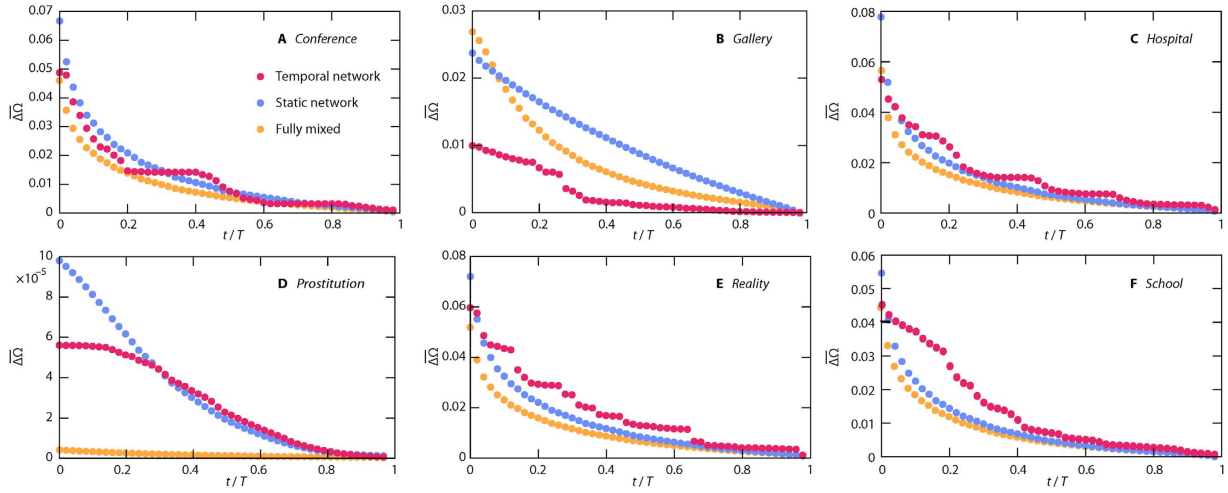

Figure 6. The time evolution of the average outbreak diversity. We investigate the average deviation of pairs of outbreak sizes (given the state of the system at time t) $\overline{\Delta\Omega}$. Here we show the results for *Gallery* (panels (**A–C**)) and *Hospital* (**D–F**). For temporal (**A,D**) and static (**B,E**) networks and a fully mixed case (**C,F**).

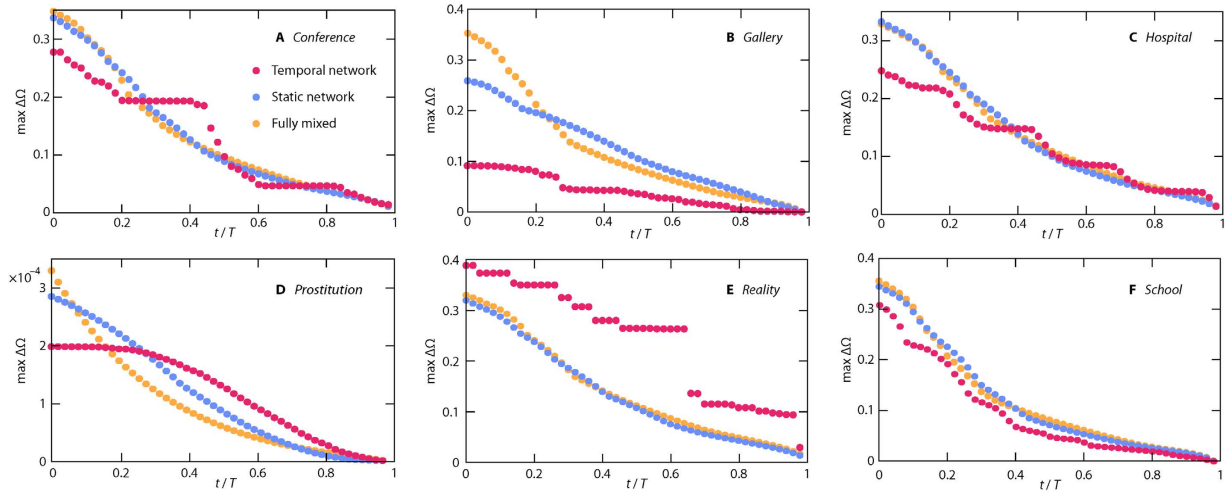

Figure 7. The time evolution of the maximum outbreak diversity. These plots are exactly corresponding to Fig. 4, but for the maximum over the parameter space, rather than the average.

Parameter dependence of time to predictability. Our final analysis regards $\Delta\Omega$'s approach to zero as a function of the SIR parameter values. In other words, we seek to summarize Fig. 3 for δ and λ at the expense of not being able to visualize the full time evolution. Instead we measure the time t_p until there is a 20-fold decrease of the outbreak diversity, i.e. when the deviation of Ω goes below 0.05 $\Omega_{\max}$, where $\Omega_{\max}$ is the Ω-value for $\delta = T$ and $\lambda = 1$. The results for the *Gallery* data are plotted in Fig. 8 (for the other data sets—see Supplementary Fig. S2). One interesting observation is that, for all the contact representations, there are parameter values where one has to wait until the very end of the sampling time to get an accurate prediction of the final outbreak size. The inherently hardest prediction happens at long durations and intermediate transmission probabilities. The fact that there is a maximum at intermediate λ is probably related to this being the region of longest outbreak times[33]—for smaller λ, the outbreak dies out when only a few individuals have been infected; for larger λ, the outbreak burns out fast in the population. Another reason for the slow approach to predictability is that the outbreaks are less deterministic[38] in this region than for larger λ (cf. the discussion of Fig. 4 above). Indeed, a large Ω does not necessarily mean a short t_p. If λ is large enough, the stochastic element disappears and the outbreak becomes predictable early (see Fig. 8C).

For the two less informative representations, the parameter-space region of slow approach to high predictability is larger. This is true for almost all the datasets (*Prostitution*, once again, being an exception,

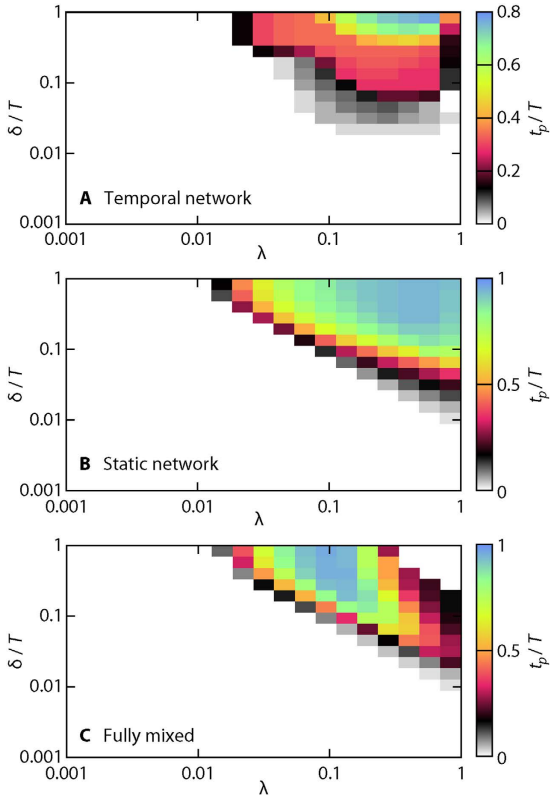

Figure 8. Time t_p to reach high predictability. We define high predictability as when the deviation of the predicted outbreak size is less than 5% of its maximal value. The data set is *Gallery*, these plots for other data sets can be found in Supplementary Fig. S2.

Supplementary Fig. S2). We also note that, there is more variation in t_p than Ω—for the example *Gallery* data of Fig. 7, all three panels have distinct shapes. For short, the parameter dependence of t_p is more complex than that of Ω. These observations holds for the other data sets with one correction—the densest static networks (*Hospital* and *Gallery*) are very similar to their fully mixed counterparts.

Discussion

We have studied how the level of information content in the representation of contacts patterns affects the SIR epidemic model. We investigated several aspects of predictability or outbreak diversity—given no knowledge about the outbreak (other than that it happened) and given the state of the system at a breaking time t. The starting point of our study was empirical data sets of human proximity. SIR outbreaks in these data sets were mostly slowed down and shrunk when a new layer of information was added (i.e. going from a fully mixed simulation to a static network representation, or going from a static network to a temporal network). Given that we do not know anything about the epidemics (more than it started), a classic (differential-equation based) analysis would overestimate the severity of the disease, as would a static-network based model. On the other hand, if we instead study the histogram of future outbreak sizes given the state of the system at time t, then there is no clear trend with respect on the information content (still, the deviations can be large). In other words, different representations do give different results, but it is, strictly speaking, not the case that adding information systematically increases or decreases the deviation of the predicted outbreak sizes. To some extent this could probably be explained as finite size effects, but higher order correlations in the temporal network could also be important. This paper only takes a first step towards understanding the relation between predictability and temporal network structure.

It is hard to generalize all features of the outbreak diversity. We note that for most data sets, including more information about the contacts makes the outbreaks smaller. However, this is not always the case (as the *Prostitution* data behaves the other way around[13,15]). Another fairly universal feature is that, for later times (initially it could be the other way around) the fully connected topologies are more predictable than the static networks. On the other hand, outbreaks on the temporal networks can be both more or less predictable. We note that the data sets with relatively short durations of the presence of nodes and links (the time between the first and last time they are observed) lose most predictability by projecting out the temporal information.

Not all our observations are vague—we can clearly see the importance of the temporal structures. Going from a temporal to a static representation can quantitatively make a big difference, but not only that—the outbreak distributions are either bi- or unimodal for the static network (and fully mixed) simulations, whereas for the temporal networks, the distribution can be multimodal (cf. Fig. 3).

This work is a starting point. To corroborate the observations above one approach would be to use generative models that can tune the temporal network structure of the data (cf. refs 10,14,39). In general, we anticipate much research relating aspects of the available information about an epidemic outbreak, and the contact structure of the population, with the predictability of the outbreak.

Methods

In this section, we will go through technicalities of the SIR model that are not fully explained above.

SIR simulations. In this work, we use the constant duration version of the SIR model[29]. We initialize all nodes to susceptible except pick one random individual i that we set as infectious. We assume that i becomes infectious at the same time as its first appearance in the data (i.e. it can infect others starting from its first contact). In a contact between an infectious and susceptible, the susceptible will (instantaneously) become infectious with a probability λ. Infectious individuals stay infectious for δ time steps whereupon they become recovered. If more than one contact occur during the same time step, we go through them in a random order.

For the fully mixed and static network models, we use the same time window of the simulation as the sampling duration of the temporal network data. We use the same number of contacts as the real data, but with the time stamps of the contacts assigned with uniform probability in the sampling window (in the static network case, they only happens between individuals connected by an edge).

Another common version of the SIR model is to let infectious individuals recover with a constant rate. Qualitatively, both versions give the same results[29]. We use the constant duration version because it both has a peaked distribution of the infection times (as opposed to the exponentially distributed times of the constant recovery rate version) and makes the code a bit faster than the exponentially distributed durations.

For all parameter values, all data sets and all representations, the output is averaged over 10^3 independent outbreaks (and every time step of every outbreak is the staring point of 10^4 auxiliary runs, as mentioned above).

References

1. Anderson, R. M. & May R. M. *Infectious diseases of humans*. (Oxford University Press, 1991).
2. Hethcote, H. W. The mathematics of infectious diseases. *SIAM Rev.* **42,** 599–653 (2000).
3. Holme, P. & Saramäki, J. Temporal networks. *Phys. Rep*. **519,** 97–125 (2012).
4. Masuda, N. & Holme, P. Predicting and controlling infectious disease epidemics using temporal networks. *F1000Prime Rep*. **5,** 6 (2013).
5. Bansal, S., Read, J., Pourbohloul, B. & Meyers L. A. The dynamic nature of contact networks in infectious disease epidemiology. *J. Biol. Dyn*. **4,** 478–489 (2010).
6. Morris, M. *Network epidemiology: A handbook for survey design and data collection*. (Oxford University Press, 2004).
7. Keeling, M. J. & Eames, K. T. D. Networks and epidemic models. *J. R. Soc. Interface* **2,** 295–307 (2005).
8. Pastor-Satorras, R., Castellano, C., van Mieghem, P. & Vespignani, A. *Epidemic processes in complex networks*. *Rev. Mod. Phys.* **87,** 925–979 (2015).
9. Holme, P. & Takaguchi, T. Time evolution of predictability of epidemics on networks. *Phys. Rev. E* **91,** 042811 (2015).
10. Holme, P. Epidemiologically optimal static networks from temporal network data. *PLoS Comput. Biol.* **9,** e1003142 (2013).
11. Liljeros, F., Edling, C. R. & Amaral, L. A. N. Sexual networks: implications for the transmission of sexually transmitted infections. *Microb. Infect.* **5,** 189–196 (2003).
12. Liu, S., Perra, N., Karsai, M. & Vespignani, A. Controlling contagion processes in activity driven networks. *Phys. Rev. Lett.* **112,** 118702 (2014).
13. Rocha, L. E. C., Liljeros, F. & Holme, P. Simulated epidemics in an empirical spatiotemporal network of 50,185 sexual contacts. *PLoS Comp. Biol.* **7,** e1001109 (2011).
14. Fefferman, N. H. & Ng, K. L. How disease models in static networks can fail to approximate disease in dynamic networks. *Phys. Rev. E* **76,** 031919 (2007).
15. Karsai, M. *et al.* Small but slow world: how network topology and burstiness slow down spreading. *Phys. Rev. E* **83,** 025102 (2011).
16. Holme, P. & Liljeros, F. Birth and death of links control disease spreading in empirical contact networks. *Sci. Rep*. **4,** 4999 (2014).
17. Lee, S., Rocha, L. E. C., Liljeros, F. & Holme, P. Exploiting temporal network structures of human interaction to effectively immunize populations. *PLoS ONE* **7,** e36439 (2012).
18. Starnini, M., Machens, A., Cattuto, C., Barrat, A. & Pastor-Satorras, R. Immunization strategies for epidemic processes in time-varying contact networks. *J. Theor. Biol.* **337,** 89–100 (2013).
19. Granell, C., Gómez, S. & Arenas, A. Dynamical interplay between awareness and epidemic spreading in multiplex networks. *Phys. Rev. Lett*. **111,** 128701 (2013).
20. Bifolchi, N., Deardon, R. & Feng, Z. Spatial approximations of network-based individual level infectious disease models. *Spatial and Spatio-temporal Epidemiology* **6,** 59–70 (2013).
21. Boccaletti, S. *et al.* The structure and dynamics of multilayer networks. *Phys. Rep*. **544,** 1–122 (2014).
22. Eagle, N. & Pentland, A. Reality mining: Sensing complex social systems. *Pers. Ubiquit. Comput*. **10,** 255–268 (2006).
23. Pfitzner, R., Scholtes, I., Garas, A., Tessone, T. J. & Schweitzer, F. Betweenness preference: Quantifying correlations in the topological dynamics of temporal networks. *Phys. Rev. Lett.* **110,** 198701 (2013).
24. Isella, L. *et al.* What's in a crowd? Analysis of face-to-face behavioral networks. *J. Theor. Biol.* **271,** 166–180 (2011).
25. Stehlé, J. *et al.* High-resolution measurements of face-to-face contact patterns in a primary school. *PLoS ONE* **6,** e23176 (2011).

26. Vanhems, P. *et al.* Estimating potential infection transmission routes in hospital wards using wearable proximity sensors. *PLoS ONE* **8,** e73970 (2013).
27. van den Broeck, W., Quaggiotto, M., Isella L., Barrat, A. & Cattuto, C. The making of sixty-nine days of close encounters at The Science Gallery. *Leonardo* **45,** 201–202 (2012).
28. Rocha, L. E. C., Liljeros, F. & Holme, P. Information dynamics shape the sexual networks of Internet-mediated prostitution. *Proc. Natl. Acad. Sci. USA* **107,** 5706–5711 (2010).
29. Holme, P. Model versions and fast algorithms for network epidemiology. *Journal of Logistical Engineering University* **30,** 1–7 (2014).
30. Bansal, S., Grenfell, B. T. & Meyers, L. A. When individual behaviour matters: homogeneous and network models in epidemiology. *J. Roy. Soc. Interface* **4,** 879–891 (2007).
31. Karsai, M., Perra, N. & Vespignani, A. Time varying networks and the weakness of strong ties. *Sci. Rep*. **4,** 4001 (2014).
32. Barthelemy, M., Barrat, A., Pastor-Satorras, R. & Vespignani, A. Velocity and hierarchical spread of epidemic outbreaks in scale-free networks. *Phys. Rev. Lett*. **92,** 178701 (2004).
33. Holme, P. Extinction times of epidemic outbreaks in networks. *PLoS ONE* **8,** e84429 (2013).
34. Meyers, L. A., Pourbohloul, B., Newman, M. E. J., Skowronski, D. M. & Brunham, R. C. Network theory and SARS: Predicting outbreak diversity. *J. Theor. Biol.* **232,** 71–81 (2005).
35. Newman, M. E. J. *Networks: An introduction*. (Oxford University Press, 2010).
36. Volz, E. M., Miller, J. C., Galvani, A. & Meyers, L. A. Effects of heterogeneous and clustered contact patterns on infectious disease dynamics. *PLoS Comput. Biol.* **7,** e1002042 (2011).
37. Holme, P. & Masuda, N. The basic reproductive number as a predictor for epidemic outbreaks in temporal networks. *PLoS ONE*. **10,** 0120567 (2015).
38. Janson, S., Luczak, M. & Windridge, P. Law of large numbers for the SIR epidemic on a random graph with given degrees. *Random Struct. Algor*. **45,** 724–761 (2013).
39. Rocha, L. E. C. & Blondel, V. D. Bursts of vertex activation and epidemics in evolving networks. *PLoS Comput. Biol*. **9,** e1002974 (2013).

Acknowledgements

This research was supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (2013R1A1A2011947). The computer simulations were carried out at the Abisko cluster of HPC2N, Umeå University.

Author Contributions

P.H. did everything.

Additional Information

Supplementary information accompanies this paper at http://www.nature.com/srep

Competing financial interests: The author declares no competing financial interests.

How to cite this article: Holme, P. Information content of contact-pattern representations and predictability of epidemic outbreaks. *Sci. Rep.* **5**, 14462; doi: 10.1038/srep14462 (2015).