Chapter 3
Sensitivity to Temporal and Topological Misinformation in Predictions of Epidemic Outbreaks

Petter Holme and Luis E.C. Rocha

Abstract Structures both in the network of who interact with whom, and the timing of these contacts, affect epidemic outbreaks. In practical applications, such information would frequently be inaccurate. In this work, we explore how the accuracy in the prediction of the final outbreak size and the time to extinction of the outbreak depend on the quality of the contact information. We find a fairly general stretched exponential dependence of the deviation from the true outbreak sizes and extinction times on the frequency of errors in both temporal and topological information.

3.1 Introduction

The propagation of infectious diseases in populations is an emergent property of the interaction between people and pathogens [1, 2]. Temporal networks is a stylized framework for describing the interaction within a population [3, 4]. It records who is in contact with whom, at what time, but omits information about the details of the encounters. In principle such details could also be important since individual, social and environmental variations affect contagion [5–7], but since our interest is to investigate the importance of temporal network structure, rather than accurate prediction, we leave them out by assuming identical individuals.

The theme of this book is to understand the role of structures in time and network topology on disease spreading. At the time of writing there are several different data sets recording the temporal contact networks of human proximity in which for example airborne diseases spread. Empirical data however is typically noisy

P. Holme (✉)
Institute of Innovative Research, Tokyo Institute of Technology, Tokyo, Japan
e-mail: holme@cns.pi.titech.ac.jp

L.E.C. Rocha
Department of Public Health Sciences, Karolinska Institutet, Stockholm, Sweden

Department of Mathematics, Université de Namur, Namur, Belgium
e-mail: luis.rocha@ki.se

N. Masuda, P. Holme (eds.), *Temporal Network Epidemiology*,
Theoretical Biology, DOI 10.1007/978-981-10-5287-3_3

either due to reporting or recording errors. This random information added to data affects the correlations in contact patterns and can potentially result in errors when analysing the data. In this chapter, we look at the sensitivity of epidemic variables with respect to simulated temporal and topological noise. Our work connects to the general questions about predictability of disease spreading [8, 9]. In this area, researchers have studied how different limitations to the models of disease spreading or incompleteness of the data affect the prediction results [10, 11]. Furthermore, researchers have investigated the internal fluctuations in the timings of contacts on the prediction of epidemic outbreaks [12–14]. The novel angle in our approach is to contrast two different types of misinformation—temporal and topological—and two different characteristics of an outbreak—the outbreak size and the time to extinction.

We use empirical, temporal proximity networks as the underlying contact structures for the disease spreading. Then we study the effects of inaccurate labeling of the nodes or time stamps on the predicted outbreak size and extinction time of susceptible-infectious-recovered (SIR) simulations on these modified data sets. The SIR model is the canonical compartmental model for diseases that gives immunity upon recovery [1]. It could be used to model e.g. HIV infection in case of treatment, that is, where the infectious individual becomes recovered after starting anti-retroviral treatment, Ebola in case of high death rate, measles and chickenpox. Simulations start with the entire population being susceptible. Then, at some point, one of the individuals becomes infectious. During this state, the infectious can spread the infection to other susceptible individuals that he or she is in contact with. As in other compartmental models, one assumes such a contagion to happen with a fixed probability per contact. After being infected for some fixed time, the infected individuals recover. When there are no infectious individuals, the outbreak is extinct. The severity of an outbreak can be quantified by various parameters. We use the outbreak size Ω (the fraction of recovered individual after the outbreak is extinct) and the extinction time τ (the duration of the outbreak in the population) as measures of the outbreak severity. In the remainder of this chapter, we will go over the background theory and technical details before we present our simulation results.

3.2 Preliminaries

3.2.1 Definitions

We represent the temporal network G as a sequence of *contacts* (i, j, t)—to be interpreted as individual i being in contact with individual j at time t [3, 4]. The number of individuals N (or *nodes*) is called the *size* of the temporal network. We use C to represent the number of contacts and M the number of pairs of individuals that are in contact at least one time. Furthermore, we let T represent the duration of the data (the time between the first and last contacts).

Table 3.1 Basic statistics of the network data sets. N is the number of individuals; C is the number of contacts; T is the total sampling time; Δt is the time resolution of the data set and M is the number of links in the projected static networks. We also list the original reference to the data

Data set	N	C	T	Δt	M	Ref.
Prostitution	16,730	50,632	6.00 y	1 d	39,044	[15]
Conference	113	20,818	2.50 d	20 s	2,196	[16]
Hospital	75	32,424	96.5 h	20 s	1,139	[17]
School	236	60,623	8.64 h	20 s	5,901	[18]
Gallery	200	5,943	7.80 h	20 s	714	[19]
Office	92	9,827	11.4 d	20 s	755	[20]

3.2.2 Contact Networks

We base our study on empirical data sets of human proximity. In other words, they capture when two persons are in close proximity, and who they are. Such data sets represent the structure over which infectious diseases spread. We list the basic statistics—sizes, sampling durations, etc.—of the data sets in Table 3.1.

Several of our data sets come from the Sociopatterns project (sociopatterns.org). These data sets are recorded by radio-frequency identification sensors that detect contacts between people within 1–1.5 m. One of these datasets comes from a conference [16] (*Conference*), another from a school (*School*) [18], a third from a hospital (*Hospital*) [17], a fourth from an art gallery (*Gallery*) [19] and a fifth from office (*Office*) [20]. The *Gallery* and *School* data sets comprise several days. We use the first day in both cases. Finally, the *Prostitution* data comes from self-reported sexual contacts between female sex-workers and their clients [15]. Since the contacts represent more than just proximity (i.e. sexual activity), this is a special form of proximity network.

3.2.3 Epidemic Simulation

The SIR simulations proceed as follows. First, all individuals are initialized to S (susceptible). Then, one node i_0 is selected randomly to become the seed of the infection. i_0 is made infectious at a random time t_0 between 0 and T. Then we go through the contacts of the data from the first to last. If the contact happens to be between a susceptible and an infectious individual, then, with a probability λ the susceptible becomes infectious. An infectious individual stays infectious δT time steps (in units of Δt) before turning recovered. In other words, δ is the fraction of the duration of the data set that a node is infectious. When there are no more infectious individuals the outbreak is considered extinct. This definition is slightly different from the most common [1], where an infectious individual has the same chance of getting recovered every time step. Our model is justified since the distribution of infectious times is narrow in real life [21] and this approach is also algorithmically simpler [22].

3.2.4 Controlling Misinformation

To model temporal misinformation, we replace a (randomly selected) fraction ϵ of the time stamps of the contacts of G by random times in the interval $[1, T]$. Similarly, for investigating the sensitivity with respect to the graph information, we randomly replace a fraction f of the node id-numbers by random numbers in the interval $[1, N]$. The only two constraints we impose in this randomization is that the resulting contacts should not be between a node and itself, and not already be present in the data. If a drawn node-id number does not satisfy the constraint, we redraw the node-id.

Technically, this approach is similar to randomization techniques [23, 24] where the temporal network structure is investigated by systematically replacing some aspect—like the timing of events—by random values and studying the response to quantities characterizing the functionality of the network (like average spreading speed, etc.). The difference is that we tune up the randomization, starting from zero (i.e. the original network).

3.2.5 Measuring Sensitivity to Misinformation

The two epidemiological variables we use to characterize an outbreak are the average final outbreak size Ω—the fraction of the population that are in state R after the outbreak is over—and the extinction time τ—the time between the first and last presence of an infected individual in the population. Let

$$\Delta_{\Omega t}(\epsilon, \delta, \lambda) = \langle \Omega(G_{\epsilon t}, \delta, \lambda) \rangle - \Omega(G, \delta, \lambda), \tag{3.1}$$

where $\langle \cdot \rangle$ denotes the average over an ensemble of networks $G_{\epsilon t}$ in which a fraction ϵ of misinformation has been imposed to the time stamps of the contacts (according to the preceding section) and G is the original network. Analogously, we define $\Delta_{\Omega n}$ for the deviation of outbreak sizes with respect to topological misinformation (i.e. rewiring of contacts generating network $G_{\epsilon n}$), and $\Delta_{\tau t}$ and $\Delta_{\tau n}$, for the deviations in the prediction of extinction times in the presence of temporal and topological noise respectively.

In principle, Δ (in any version) could be negative, but for our data sets that rarely happens—the practical minimum is $\Delta = 0$ for $\lambda = 0$. To study the ϵ dependence of Δ, we need to look at a summary statistic over the SIR parameter space. In this work, we will focus on the worst case scenario. We will use the summary statistic

$$\omega(\epsilon) = \max_{\delta,\lambda} \Delta(\epsilon, \delta, \lambda) - \min_{\delta,\lambda} \Delta(\epsilon, \delta, \lambda). \tag{3.2}$$

I.e. the difference in the range of Δ values. This quantity will be dominated by $\max_{\delta,\lambda} \Delta$, but also give a slight extra weight to networks with a negative $\min_{\delta,\lambda} \Delta$.

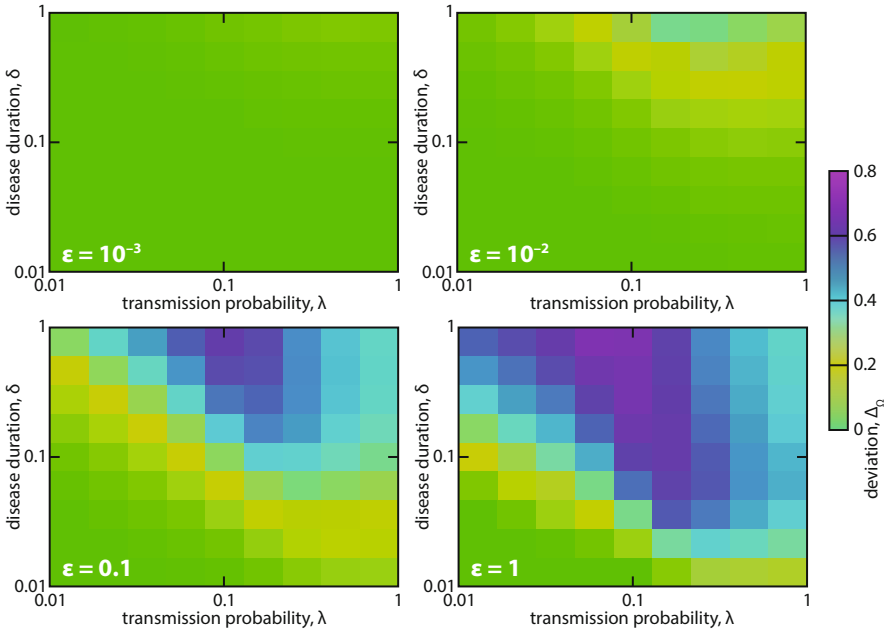

Fig. 3.1 Heatmap of the difference $\Delta\Omega$ between the average outbreak size Ω for the *Hospital* manipulated data set, where we vary a fraction ϵ of misinformation in the node identities, and the original data set. The different panels represent different values of the error rate ϵ

3.3 Results

3.3.1 Impact of Misinformation Throughout the SIR Parameter Space

As a first numerical study, we investigate $\Delta_{\Omega n}(\epsilon, \delta, \lambda)$ (Fig. 3.1) and $\Delta_{\tau n}(\epsilon, \delta, \lambda)$ (Fig. 3.2) for the *Hospital* data set. We chose this data set as a case study because it is of intermediate size and heterogeneity both in the temporal and topological structure. It is also highly relevant for the spread of healthcare associated infections [25]. We study an exponential sequence of ϵ-values—$\epsilon = 10^{-3}, 10^{-2}, 10^{-1}, 1$—and, in the first place, only misinformation concerning the node identities. As seen in Fig. 3.1, the response to the noise is a non-linear function of both ϵ, δ and λ. For $\epsilon = 10^{-4}$, the impact is less than $\Delta_{\Omega n} < 0.1$ throughout the SIR parameter space. For $\epsilon = 10^{-3}$, it reaches values around 0.2, while for larger ϵ-values, $\Delta_{\Omega n} > 0.5$ for a large part of the parameter space. The shape of the region of high deviation also changes with ϵ. It seems, rather universally, the case that Δ reaches its maximum for large δ-values, but for large ϵ, also relatively small δ-values can show large deviations.

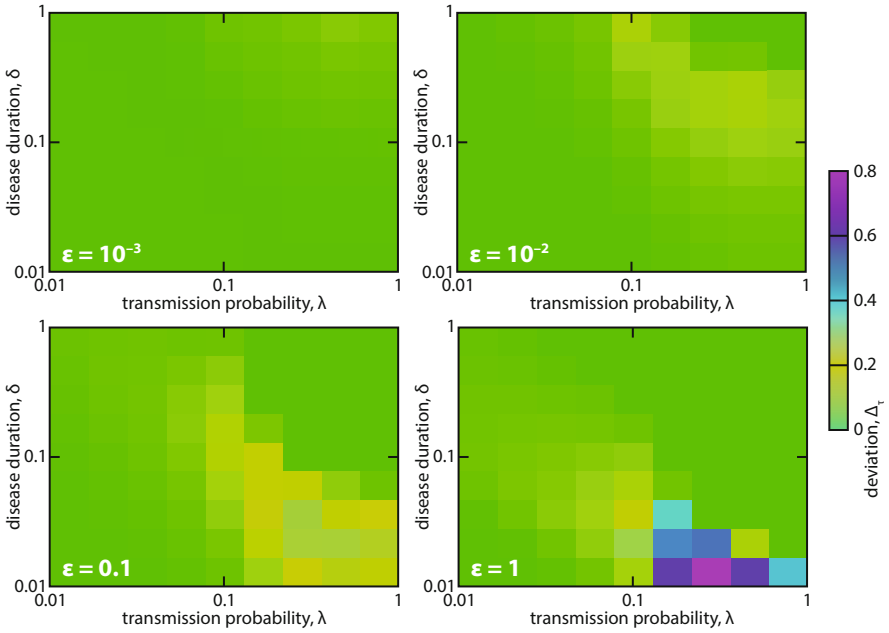

Fig. 3.2 Heat map of the difference $\Delta\tau$ between the average time to extinction τ for the *Hospital* manipulated data set, where we vary a fraction ϵ of misinformation in the node identities, and the original data set. The different panels represent different values of the error rate ϵ

For prediction of the extinction times, the absolute values of Δ are smaller for temporal misinformation in comparison to topological misinformation (Fig. 3.2). In other words, in the worst case, the prediction is somewhat better for τ than Ω. Furthermore, the parameter dependence is quite different. The maximal deviations happen for small δ-values. This is not so surprising—for relatively high values of δ and λ, the outbreak will last as long as the length T of the data set, thus making Δ small. If δ is small enough, the disease will die out without spreading much and thus Δ will also be rather small for small δ-values—the largest Δ_τ thus happens for intermediate δ.

The impact of temporal misinformation follows a similar picture to the impact of topological misinformation. The average outbreak sizes Ω differs most from the unperturbed network when the disease duration is as long as possible (Fig. 3.3). The impact changes non-linearly with both δ and λ. For the extinction time (Fig. 3.4), the situation is a bit different however. Now the largest impact does not necessarily happen for the largest δ-values. Whereas for $\epsilon \leq 10^{-2}$ it does happen at the largest δ, when $\epsilon > 10^{-2}$ the maximum is located at lower δ-values.

Several of the observations for the *Hospital* network holds for other data sets as well. However, the *Prostitution* network has a fairly different pattern (with negative $\Delta_{\Omega t}$ values for a large part of the parameter space). The origin of this anomalous behavior comes from the growth of the data (the number of contacts per time unit and the number of individuals present) was first pointed out in Ref. [14] and discussed further in Ref. [26].

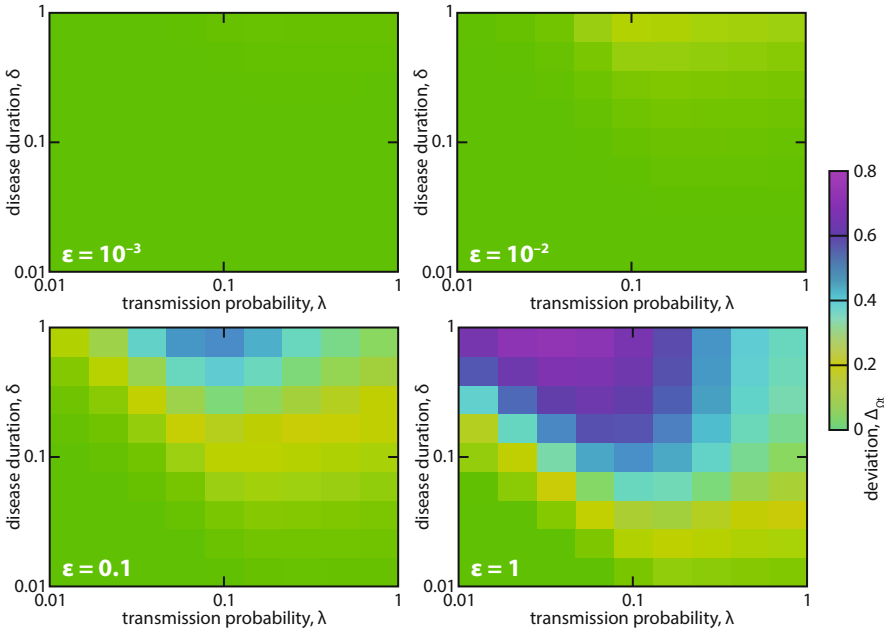

Fig. 3.3 Heat map corresponding to Fig. 3.1, but for misinformation about the timing of contacts

3.3.2 Impact of Error Rate on Prediction

To better understand the response of the level of misinformation on the prediction accuracy, we study $\omega(\epsilon)$—the difference between the largest and smallest Δ-values (Eq. 3.2). The results for this quantity are displayed in Figs. 3.5 (for Ω) and 3.6 (for τ). The lower limit of ω is trivially $\omega(\epsilon = 0) = 0$. The shape of the $\omega(\epsilon)$ is concave (meaning the effect of increasing ϵ is largest for small ϵ). In fact, we find the functional form fitting well to a stretched exponential convergence

$$\widetilde{\omega}(\epsilon) = \omega_{\max}\left[1 - \exp(-a\epsilon^b)\right], \tag{3.3}$$

where a and b are fitting parameters. The parameter b (typically in the interval $0 < b < 1$) is called the *stretching exponent* and its deviation from one indicates how much the tail is stretched compared to an exponential decay [27]. As far as we can see, there is no simple explanation for this functional form. Rather, we believe that in general the $\omega(\epsilon)$-curves can have other shapes than stretched exponentials. Indeed, the points that are off the fitting curves (e.g. the second point in the *Gallery* graph of Fig. 3.6) are probably not a result of bad convergence, but structures in the data sets. The three fitting parameters of Eq. 3.3 are nevertheless concise ways of summarizing the shapes of the $\omega(\epsilon)$-curves and revealing how the temporal network structure influences the impact of misinformation.

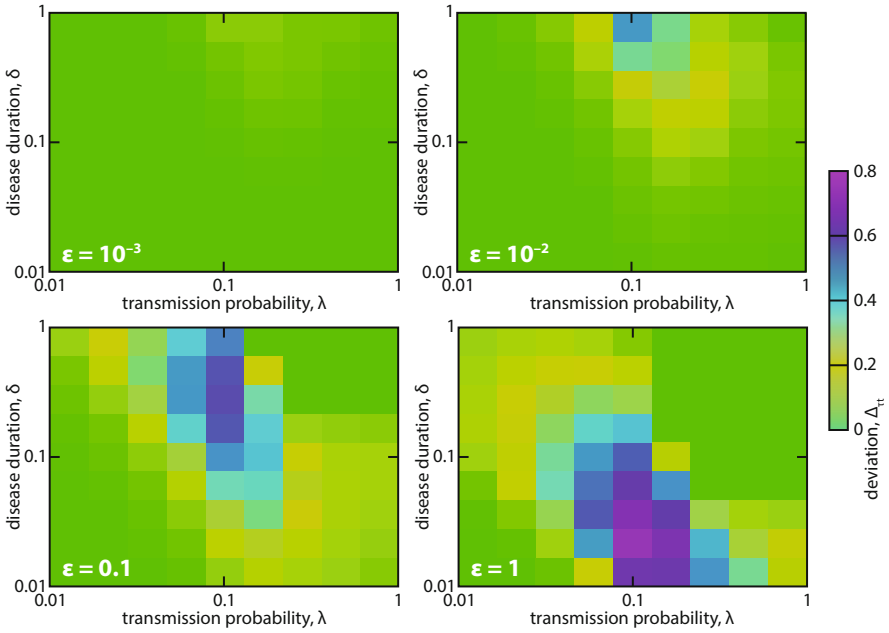

Fig. 3.4 Heat map corresponding to Fig. 3.2, but for misinformation about the timing of contacts

As alluded to, the perhaps most interesting parameter of the stretched exponential fits is the stretching exponent b. If $b = 1$, the decay is exponential. If $b < 1$ the decay is stretched (or slower than exponential). For a given error rate ϵ, the difference Δ is larger for small b. As seen in Fig. 3.7, it is indeed the case for all scenarios and data sets that $0 < b < 1$. The sparsest data set (in terms of number of contacts per individual), *Prostitution*, has a stretching exponent close to one. For the other data sets there is at least one exponent that is far off from one. There is, however, no straightforward explanation for the values of the stretching exponents in terms of the basic parameters of the temporal network data sets (as listed in Table 3.1). In future work, we will seek explanations in terms of quantities describing the temporal network structure [26, 28]. The smaller values of b for the *Conference* and *Hospital* data sets in case of τ_t happen because if we redistribute the time stamps, there will be less chance for the epidemics to die in comparison to the original data in which contacts are more concentrated at certain intervals of time. Note that in both cases, we consider night periods that correspond to absence of activity in the original data set. For the *School* data set, where individuals are clustered into network communities (i.e. the classes) the outbreak in the manipulated network is much larger since a weak randomization of id-numbers is sufficient to better distribute the links, making the network more random and thus facilitating the disease spread to the entire network. Note that in this case, there are many links at a given time step and thus the distribution of time stamps will not be much affected.

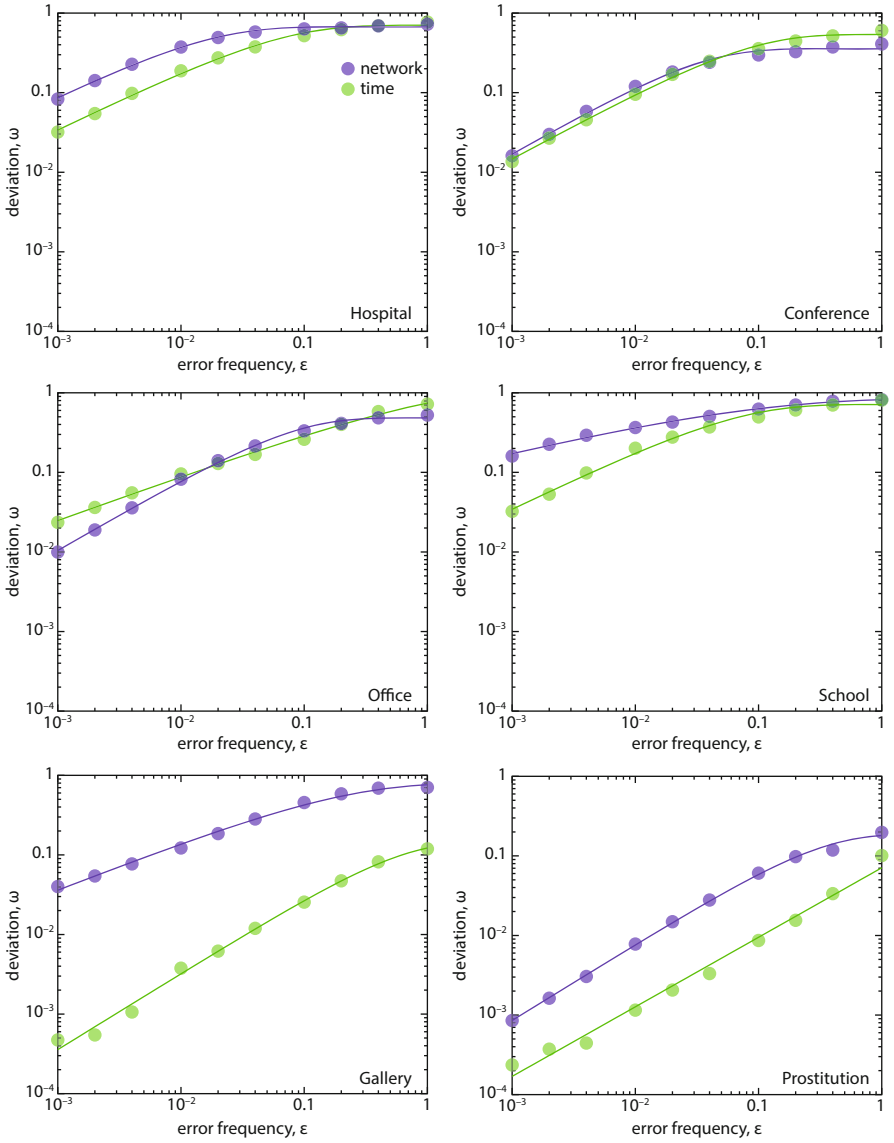

Fig. 3.5 ω_Ω, the difference between the largest and smallest Δ values over the SIR parameter space as a function of the node-identity misinformation frequency ϵ. The *curves* are Levenberg–Marquardt fits to a stretched exponential form, $\omega_{\max}(1 - e^{-a\epsilon^b})$

Gallery is a special case because groups of individuals visit the museum at fixed time slots. Possibly in this case, the disease spreads for longer times after redistributing the nodes because new links are now made between early and late museum visitors. This effect may sustain the disease for longer times and also affect the outbreak size.

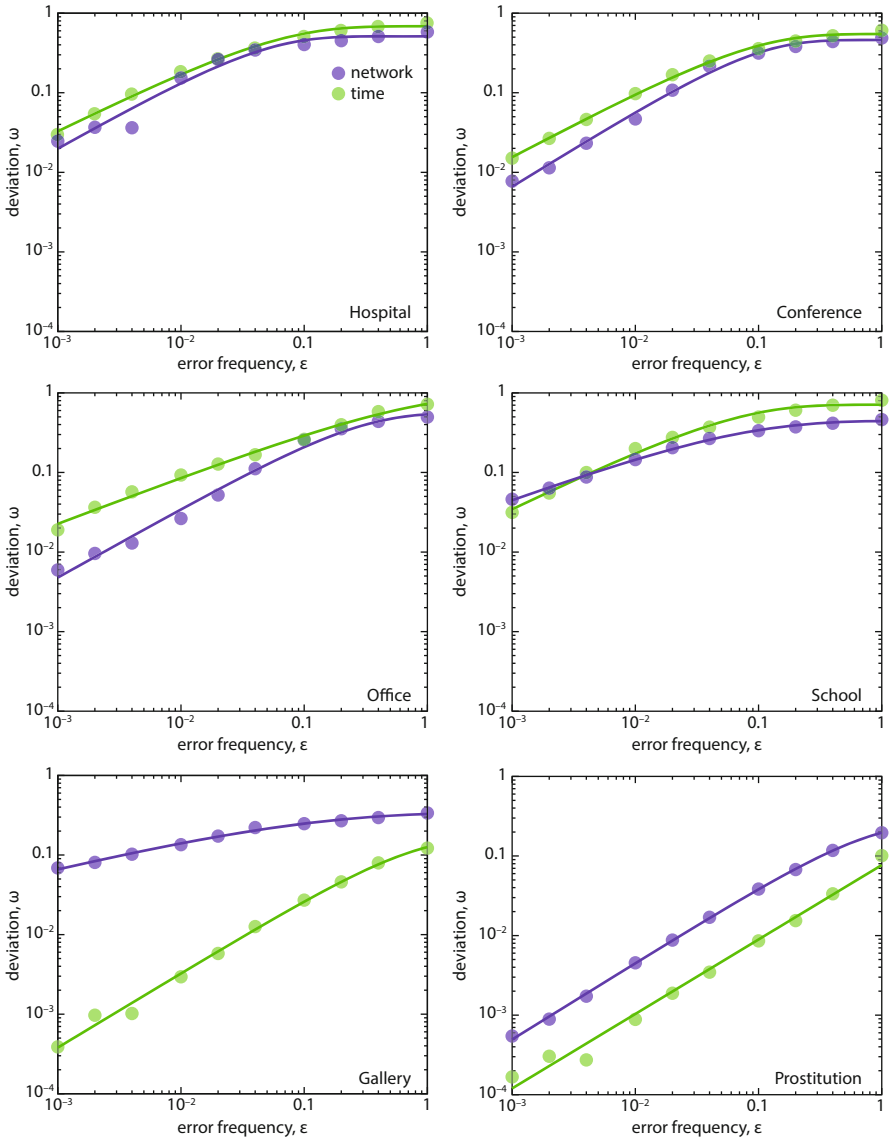

Fig. 3.6 The figure corresponding to Fig. 3.5 but for extinction times τ rather than outbreak sizes Ω

However, redistributing the time stamps will have little impact since individuals were not uniformly active during the day, for example, a new time stamp may occur at time $t = 10$ for an originally late visitor, i.e. all other connections are made at later times, therefore this new link does not contribute much to the disease spread.

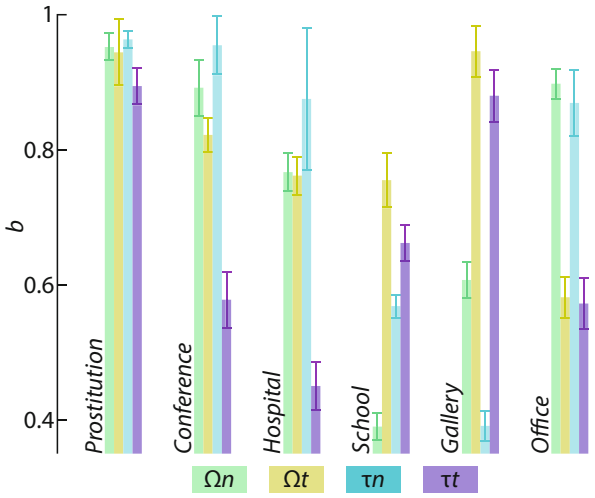

Fig. 3.7 The the exponent b of the stretched exponential fits in Figs. 3.5 and 3.6. As elsewhere, Ωn relates to the response to the prediction of the average outbreak size Ω in the case of misinformation in the node identity information. Error bars represent standard errors

3.4 Discussion

In this work, we have investigated the ability to predict outbreaks of disease given imperfect data on the temporal contacts of a population. We contrast misinformation in the identities of the individuals and the time stamps of the contacts. For both misinformation scenarios, the deviation from the accurate prediction can reach 80% for 100% error frequency ϵ. Even for small errors, the deviation may differ 10 to 20% for some epidemiological parameters. However, the area in the parameter space of such a bad prediction is rather small. Furthermore, the functional dependency of the degree of mis-prediction on ϵ is similar for the two scenarios—a stretched exponential decay. At this point, we do not have any explanation for this behavior. It would be interesting to know the conditions on the temporal network structure for such a stretched exponential decay to occur.

In a wider context, this work further illustrates the importance of temporal structure for predicting disease spreading—it seems as important as the topological information. This is along the lines of observations in e.g. Refs. [26, 29–31], and a further reason for theoretical epidemiologists to investigate the role of the temporal structure in human contact patterns for disease spreading.

It would be interesting to explore this problem with alternative models for the misinformation. In real contact patterns, there would probably be more missing contacts [11] than false contacts—i.e. the assumption that the number of contacts is preserved as ϵ increases could probably be relaxed. Another step towards increased realism would be to assume the time stamps deviate from their true value by some random variable. This is expected in data collection surveys where participants have to remember the dates of events, for example, of sexual contacts [32], or when the date of the event is recorded at random times after the actual date of the event [15]. There are many other directions to proceed towards an understanding of the relation of incomplete information and the prediction of epidemics.

References

1. Hethcote, H.W.: Infectious diseases in humans. SIAM Rev. **32**(4), 599 (2000)
2. Keeling, M.J., Eames, K.T.: Social factors in epidemiology. J. R. Soc. Interface **2**(4), 295 (2005)
3. Holme, P., Saramäki, J.: Nosocomial infections. Phys. Rep. **519**(3), 97 (2012)
4. Holme, P.: The making of sixty nine days of close encounters at the science gallery. Eur. Phys. J. B **88**, 234 (2015)
5. Giesecke, J.: Modern Infectious Disease Epidemiology, 2nd edn. Arnold, London (2002)
6. Anderson, R.M., May, R.M.: Infectious Diseases in Humans. Oxford University Press, Oxford (1992)
7. Bauch, C.T., Galvani, A.P.: Compensating for population sampling in simulations of epidemic spread on temporal contact networks. Science **32**, 47 (2013)
8. Colizza, V., Barrat, A., Barthélemy, M., Vespignani, A.: Modern infectious disease epidemiology. BMC Med. **5**(1), 34 (2007)
9. Holme, P.: Data on face-to-face contacts in an office building suggest a low-cost vaccination strategy based on community linkers. Sci. Rep. **5**, 14462 (2015)
10. Dawson, P.M., Werkman, M., Brooks-Pollock, E., Tildesley, M.J.: The mathematics of infectious diseases. Proc. R. Soc. Lond. B Biol. Sci. **282**(1808), 20150205 (2015)
11. Génois, M., Vestergaard, C.L., Cattuto, C., Barrat, A.: Social organization patterns can lower disease risk without associated disease avoidance or immunity. Nat. Commun. **6**, 8860 EP (2015)
12. Meyers, L.A., Pourbohloul, B., Newman, M., Skowronski, D.M., Brunham, R.C.: Network reachability of real-world contact sequences. J. Theor. Biol. **232**, 7181 (2005)
13. Rocha, L.E.C., Blondel, V.D.: Model versions and fast algorithms for network epidemiology. PLoS Comput. Biol. **9**(3), e1002974 (2013)
14. Rocha, L.E.C., Liljeros, F., Holme, P.: Information content of contact-pattern representations and predictability of epidemic outbreaks. PLoS Comput. Biol. **7**(3), e1001109 (2011)
15. Rocha, L.E.C., Liljeros, F., Holme, P.: Modern temporal network theory: a colloquium. Proc. Natl. Acad. Sci. USA **107**, 5706 (2010)
16. Isella, L., Stehlé, J., Barrat, A., Cattuto, C., Pinton, J.F., van den Broeck, W.: Temporal network structures controlling disease spreading. J. Theor. Biol. **271**, 166 (2011)
17. Vanhems, P., Barrat, A., Cattuto, C., Pinton, J.F., Khanafer, N., Régis, C., Kim, B.A., Comte, B., Voirin, N.: Birth and death of links control disease spreading in empirical contact networks. PLoS One **8**, e73970 (2013)
18. Stehlé, J., Voirin, N., Barrat, A., Cattuto, C., Isella, L., Pinton, J.F., Quaggiotto, M., van den Broeck, W., Régis, C., Lina, B., Vanhems, P.: The basic reproduction number as a predictor for epidemic outbreaks in temporal networks. PLoS One **6**, e23176 (2011)
19. Van den Broeck, W., Quaggiotto, M., Isella, L., Barrat, A., Cattuto, C.: Temporal networks. Leonardo **45**(3), 285 (2012)
20. Génois, M., Vestergaard, C.L., Fournet, J., Panisson, A., Bonmarin, I., Barrat, A.: What's in a crowd? Analysis of face-to-face behavioral networks. Netw. Sci. **3**, 326 (2015)
21. Lloyd, A.L.: Small but slow world: how network topology and burstiness slow down spreading. Theor. Popul. Biol. **60**, 59 (2001)
22. Holme, P.: J. Networks and epidemic models. Logist. Eng. Univ. **5**, 51 (2014)
23. Holme, P.: Stretched exponential distributions in nature and economy: "fat tails" with characteristic scales. Phys. Rev. E **71**(4), 046119 (2005)
24. Karsai, M., Kivelä, M., Pan, R.K., Kaski, K., Kertész, J., Barabási, A.L., Saramäki, J.: Realistic distributions of infectious periods in epidemic models: changing patterns of persistence and dynamics. Phys. Rev. E **83**(2), 025102R (2011)
25. Breathnach, A.S.: Network theory and SARS: predicting outbreak diversity. Medicine **33**(3), 22 (2005)

26. Holme, P.: Bursts of vertex activation and epidemics in evolving networks. Phys. Rev. E **64**, 022305 (2016)
27. Laherrère, J., Information dynamics shape the sexual networks of internet-mediated prostitution. Sornette, D.: Eur. Phys. J. B **2**(4), 525 (1998)
28. Holme, P., Masuda, N.: Simulated epidemics in an empirical spatiotemporal network of 50,185 sexual contacts. PLoS One **10**(3), e0120567 (2015)
29. Hock, K., Fefferman, N.H.: High-resolution measurements of face-to-face contact patterns in a primary school. Ecol. Complex. **12**, 34 (2012)
30. Holme, P., Liljeros, F.: Implementation of web-based respondent driven sampling among men who have sex with men in Sweden. Sci. Rep. **4**, 4999 (2014)
31. Valdano, E., Poletto, C., Giovannini, A., Palma, D., Savini, L., Colizza, V.: Predicting epidemic risk from past temporal contact data. PLoS Comput. Biol. **11**(3), e1004152 (2015)
32. Strömdahl, S., Lu, X., Bengtsson, L., Liljeros, F., Thorson, A.: Estimating potential infection transmission routes in hospital wards using wearable proximity sensors. PLoS One **10**(10), e0138599 (2015)