An algorithm to build synthetic temporal contact networks based on close-proximity interactions data

Short title: An algorithm to build synthetic contact networks from contact data

Audrey Duval[1,2,3+], Quentin Leclerc[1,2,4+], Didier Guillemot[1,2], Laura Temime[4,5#], Lulla Opatowski[1,2#]

[+] these authors contributed equally

[#] these authors contributed equally

[1]Epidemiology and Modelling of Bacterial Escape to Antimicrobials, Department of Global Health (EMEA), Université Paris Cité, Institut Pasteur, Paris, France

[2]Echappement aux Anti-infectieux et Pharmacoépidémiologie U1018, CESP, INSERM, Université Paris-Saclay, Université de Versailles St-Quentin-en-Yvelines, France

[3]Imagine Institute, Data Science Platform, INSERM UMR 1163, Université de Paris, Paris, France

[4]Laboratoire Modélisation, Epidémiologie et Surveillance des Risques Sanitaires (MESuRS), Conservatoire National des Arts et Métiers, Paris, France

[5]Unité PACRI, Institut Pasteur, Conservatoire National des Arts et Métiers, Paris, France

Keywords: long-term care facility, contact network, close-proximity interactions, sensors, network reconstruction

Acknowledgments

AD, LT and LO received funding from the French National Research Agency (SPHINX-17-CE36-0008-01).

Abstract

Small populations (e.g., hospitals, schools or workplaces) are characterised by high contact heterogeneity and stochasticity affecting pathogen transmission dynamics. The increased availability of empirical individual contact data provides unprecedented information to characterize such heterogeneity. However, these detailed data are usually collected over a limited period, and can suffer from observation bias. We propose an algorithm to stochastically reconstruct realistic temporal networks from individual contact data in health care settings (HCS) and test this approach using real data previously collected in a long-term care facility (LTCF).

Our algorithm generates full networks from recorded close-proximity interactions, using hourly inter-individual contact rates and information on individuals' wards, the categories of staff involved in contacts, and the frequency of recurring contacts. It also provides data augmentation by reconstructing contacts for days when some individuals are present in the HCS without having contacts recorded in the empirical data. Recoding bias is formalized through an observation model, to allow direct comparison between the augmented and observed networks.

The algorithm successfully reconstructed unobserved contacts, and was substantially more accurate to reproduce network characteristics than random graphs. The reconstructed networks reproduced well the assortativity by ward (first–third quartiles observed: 0.54–0.64; synthetic: 0.52–0.64) and the hourly staff and patient contact patterns. Importantly, the observed temporal correlation was also well reproduced (0.39–0.50 vs 0.37–0.44), indicating that our algorithm could recreate a realistic temporal structure. The daily degree (10.8–14.7 vs 14.7–18.0), density (0.07–0.08 vs 0.08–0.10) and global efficiency (0.39–0.43; 0.46–0.49) were slightly overestimated.

To conclude, we propose an approach to generate realistic temporal contact networks and reconstruct unobserved contacts from summary statistics computed using individual-level interaction networks. This could be applied and extended to generate contact networks to other settings using limited empirical data, to subsequently inform individual-based epidemic models.

Introduction

Limiting the public health burden of infectious diseases requires a good understanding of how they spread. For diseases transmitted mostly via close-proximity interactions, the rate at which individuals come into contact with each other is strongly correlated with the expected spread of the disease across the population [1]. In large populations such as cities or countries, contact structures can be approximated by grouping individuals into relatively broad categories (neighbourhood, age…), and assuming that contact rates are heterogeneous between categories, but homogeneous within [2,3]. In small populations such as healthcare institutions, schools, or workplaces however, disease transmission is affected by high contact heterogeneity and stochasticity [4]. Capturing these characteristics requires a detailed, individual-level description of contacts.

Contact networks are increasingly used to fully capture the interactions between individuals in small populations [5,6]. These networks explicitly represent the links between all individuals in such populations, as opposed to contact matrices which only capture average contact rates between groups of individuals [7,8]. Temporal contact networks further capture the time-changing nature of contacts, therefore representing individual interactions more accurately than static networks [9–13]. Contact networks can be coupled with individual-based mathematical models to help design effective interventions against the spread of infectious diseases, since they enable the identification of highly connected individuals who can be targeted to lead to the greatest impact on transmission [8]. Recently, empirical data collected to build inter-individual temporal networks has become increasingly available. For example, previous studies have used sensors to record close-proximity interactions between individuals [14,15], and contact tracing programs have relied on the integrated Bluetooth technology in mobile phones [16].

However, the detailed empirical data required to build temporal contact networks remain subject to several limitations [17,18]. Even studies designed to collect such data are usually limited in time, and may be subject to observation bias; sensors might not be properly placed to register contacts [19], or individuals may disable Bluetooth on their mobile phones at different times [16]. Due to the resulting missed contacts, the networks derived from these data may only be partially observed. Transmission rates estimated using these partially

observed networks would be overestimated compared to reality due to the lower number of contacts, which could lead to an incorrect evaluation of the impact of interventions [20–22]. By comparison, although they do not provide individual-level information, contact matrices and summary statistics such as contact rates between individual groups are more readily available, as they can be inferred using simple cross-sectional survey data [23–25].

Here, we propose an algorithm to stochastically reconstruct realistic contact networks from partially observed contact data. To validate this approach, we use close-proximity data collected in a long-term care facility (LTCF) during the i-Bird study [1]. We compute summary contact parameters from these data to generate reconstructed contact networks and compare these synthetic contact networks with the observed i-Bird data.

Methods

Data description

The data used here were collected during the Individual-Based Investigation of Resistance Dissemination (i-Bird) study [1]. This study took place in a rehabilitation and long-term care facility (LTCF) from the beginning of July to the end of October 2009. Over this period, each participant (patient or hospital staff) was wearing an RFID sensor that recorded close-proximity interactions (CPIs, at less than 1.5m) every 30 seconds. Here, we only used contacts recorded between 27 July to 23 August 2009 (included). This period corresponds to the weeks between two sensor battery replacements and hence avoids interference due to loss of contact. A temporal network of proximities was therefore available over 28 days with information on individual ID and ward of affectation.

The hospital was structured into five wards: three neurological wards, one nutritional care ward and one geriatric ward. Patients were systematically linked to a ward, whilst some staff were mobile and not linked to a specific ward. For the purpose of this work, we consider here that mobile staff belong to an "artificial" 6^{th} ward, to compute contact rates according to the algorithm detailed below. Staff were divided into 13 professions: administrative, animation/hairdresser, logistic, hospital service agent, porter, occupational therapist, physiotherapist, other re-education, nurse, head nurse, care assistant, student/intern, and physician. A total of 200 patients and 213 hospital staff were included and had contacts recorded during the 28 days of study.

We used hospital staff schedules to determine the hourly presence of each staff. We compared these schedules, as well as admission and discharge dates of patients, to the dates and times when individuals had any contact recorded. Through this, we estimated that there was on average no contact data recorded for 37% (standard deviation: 30%) of a patient's presence days, and 42% (sd: 38%) for staff (Supplementary Figure 1). The raw i-Bird network, measured directly from the sensors, is therefore an incomplete representation of the real inter-individual proximity network over the period and underestimates the number of contacts in the hospital. Interestingly, at the population level, there was no correlation between the proportion of presence time during which contact data were recorded for a given

individual and their average number of contacts on presence days where data were available (Supplementary Figure 2).

Building synthetic contacts

Algorithm outline

We built an algorithm to stochastically reconstruct a realistic full temporal network of inter-individual close-proximity interactions in the hospital using parameters estimated from the observed i-Bird data. This algorithm generates a new synthetic network which notably reconstructs contacts over days when individuals were known to be present in the hospital but had no contact data recorded, which we consider to be a recording bias. The synthetic network hence includes both the observed and unobserved parts of the empiric network. This approach first involves the calculation of contact rates and durations between individuals, stratified by the individuals' ward, category (patient, or staff profession), type of day (weekday or weekend) and hour. The algorithm then reconstructs a new network, taking as input these summary statistics as well as data on presence days for each individual in the facility. Each CPI is generated stochastically, with individuals chosen in order to promote recurring contacts, based on a probability estimated from the data.

Estimation of contact rates from the i-Bird data

Contact rates per hour (h from 00h to 23h), category of individual (C_i, i.e. patient, or hospital staff profession) and ward W_i are estimated from the data as:

$$T_{h,c_1w_1\to c_2w_2} = \frac{\sum_{i\in C_1W_1}\sum_{j\in C_2W_2}\sum_{k=1}^{N_{h,i}} V_{i,j,k}}{\sum_{l=1}^{N_h} N_{C_1W_1,l}} \qquad (1)$$

where $T_{h,c1w1\to c2w2}$ is the average per-person contact rate at the hour h between individuals from category C_1 belonging to the ward W_1 and individuals from category C_2 belonging to the ward W_2. For given hour h and individual i, $N_{h,i}$ is the number of instances of the hour h where at least one contact was recorded for that individual. For example, if i had a contact recorded on Tuesday 11th August at 10h, and on Tuesday 18th August at 10h, $N_{10,i}$ would be equal to 2. For two individuals i from C1W1 and j from C2W2, $V_{i,j,k}$ indicates whether contacts have been recorded on occurrence k: it equals 1 if i and j had at least one contact recorded at the instance

k of the hour *h*, and 0 otherwise. Finally, N_h is the total number of instances of the hour *h* in the full dataset and, for a given instance *I* of the hour *h*, $N_{C1W1,I}$ is the number of individuals from C_1W_1 that had any contact recorded during that hour.

This estimation was conducted separately for contacts during weekdays and contacts during weekends.

Estimation of recurring contacts

For each individual *i*, we calculate the probability of recurring contact for each day *d* between the first (d_0) and last (d_{max}) days where a contact was recorded for *i*, according to

$$p_{i,d} = \frac{|U_{i,d} \cap U_{i,[d_0,d]}|}{|U_{i,d}|} \qquad (2)$$

Where $U_{i,d}$ is the set of unique individuals with whom *i* had a contact on day *d*, and $U_{i,[d0,d]}$ is the set of unique individuals with whom *i* had at least one contact on any day between the first day d_0 and the current day *d*. For example, if *i* had a contact with four unique individuals on day *d*, and previously had a contact with two of those on any day between d_0 and *d*, the probability of recurring contact for day $p_{i,d}$ would be 2/4 = 0.5.

We then calculated the mean probability of recurring contacts for individual *i* across all days as

$$p_i = \frac{\sum_{d=d_0}^{dmax} p_{i,d}}{1+(d_{max}-d_0)} \qquad (3)$$

Finally, we calculated the mean probability of recurring contacts by individual category *c* (patient or staff) as

$$p_c = \frac{\sum_{i \in C} p_i}{|C|} \qquad (4)$$

Where *C* represents the set of individuals belonging to category *c*.

Generation of synthetic CPIs: number and individuals in contacts

For each hour of our period of interest, we estimate the number of contacts between individuals present in the hospital during that hour, determined using the admission data and staff schedule collected during the i-Bird study. We generate the number of individuals *n* from category C_2S_2 in contact with an individual *i* from category C_1S_1 during an hour *h* by sampling

from a Poisson distribution with the mean being the contact rate as described above. Before selecting these *n* individuals, since contacts are generated dynamically, we check if *i* is already included in the contacts of individuals from C_2S_2 during *h*. If *n'* individuals from C_2S_2 have already had a contact with *i* during *h*, we only select *n = n-n'* new individuals from those available, in order to avoid double counting.

These *n* individuals are selected by favouring contacts between individuals who have already met at any other time previous to *h*. Let p_c be the probability of a recurring contact for category *c* (patient or staff) of the individual *i*. To determine the identity of the *n* individuals in contact with *i*, we draw a random number $r \sim Uniform(0,1)$

- If $r \leq p_c$, a recurring contact is generated: *j* is chosen among *S*, the subset of C_2S_2 individuals who previously met *i*, according to probability $p_{i \to j}$:

$$p_{i \to j} = \frac{N_{i \to j}}{\sum_{k \in S} N_{i \to k}} \qquad (5)$$

  Where $N_{i \to j}$ is the number of previous contacts between *i* and *j* before hour *h*, and $N_{i \to k}$ is the number of previous contacts between *i* and each individual *k* belonging to *S*.
- Otherwise, the contact is not recurring: the individual *j* in contact is randomly and uniformly chosen among *S'*, the subset of C_2S_2 individuals who have not yet met *i*.

Generation of contact durations

We generated the duration of contact between two individuals (*i* from C_1S_1 and *j* from C_2S_2) by sampling from a log-normal distribution. This distribution was calibrated using the mean and variance of the duration of contact at the hour *h* between two individuals from C_1S_1 and C_2S_2, estimated from the data.

Observation bias process

As mentioned earlier, in any real-life data, it can be assumed that there are periods of non-recording of CPIs (bias in collection). For each individual in the observed network, we identified the hours when they had no contact recorded. We then removed those individuals on those hours before proceeding with the algorithm described above. Hence, the reconstructed biased network and the observed network suffer from the same bias and are comparable.

Simulations and analysis

From the observed network, we generated 100 full reconstructed networks, and 100 reconstructed networks with observation bias. For comparison, we also generated 100 pseudo-random contact networks with observation bias, and 100 without. The latter networks simulate contacts without taking into account the ward, staff category, and probability of recurring contact in the calculation of contact rates and durations. At each contact, the individual encountered is therefore chosen randomly from all those present in the hospital at that time, regardless of whether or not the individual was previously encountered.

We implemented the algorithm in C++ with the repast HPC 2.3.0 library. All simulations were performed on the Maestro cluster hosted by the Institut Pasteur. The networks were analysed in R [27], using the igraph package [28]. The relevant contact networks and analysis code are available in the following GitHub repository: https://github.com/qleclerc/network_algorithm.

Validation of the full reconstructed networks

For validation, we also applied the algorithm to each of the 100 reconstructed networks with bias, to generate 100 new full reconstructed networks and confirm these “re-simulated networks” were similar to the full reconstructed networks generated from the observed data.

Results

Description of the observed network, application to the i-Bird dataset

In this section, the contact data are aggregated at the daily level, so that if two individuals have two separate contacts with each other at different times of the day, this is only counted once. The contact network is considered undirected, since contacts are assumed to be reciprocal. Daily-averaged contact matrices built from these data are described in a previous work [1].

We first summarise the observed temporal network, comparing the total network and subgraphs with only patient-patient, staff-staff, or patient-staff contacts (Figure 1a-d). Table 1 provides the degree, global efficiency, density, transitivity, assortativity and temporal correlation of these four networks. The mean degree of the total network per day is 12.99 (standard deviation: 3.53), which corresponds to the average number of unique contacts per individual per day. In the subgraphs, the degree is highest in the patient-staff subgraph (8.09; sd: 1.89), although we still note a relatively important number of patient-patient contacts, with a degree of 5.25 (sd: 1.87) in the corresponding subgraph. The distribution of individual degrees for all individuals and all days across the total network is heterogeneous, with a squared coefficient of variation equal to 0.44 (Figure 1e). The global efficiency of the total network is 0.40 (sd: 0.05), meaning that on average the shortest path between any two individuals has a distance of 2.5 (whereby the shortest path between two individuals in direct contact would be of distance 1). As expected, the efficiencies are lower in the subgraphs, since we remove individuals and hence increase the distance between those remaining (patient-patient: 0.25 (sd: 0.08), staff-staff: 0.32 (sd: 0.10), patient-staff: 0.31 (sd: 0.05)). Densities in the total network and subgraphs are relatively low (< 0.1), indicating that less than 10% of all possible connections between individuals in the network are actual observed connections.

Transitivity in the total network is high (0.37; sd: 0.02), meaning that for any two individuals *a* and *b* both in contact with the same third individual *c*, the probability that *a* and *b* are also in contact is 0.37. Transitivity is also high in the patient-patient and staff-staff subgraphs, but this metric is not relevant for the patient-staff subgraph – it is impossible for a triangle of

contacts to occur in this subgraph as it excludes staff-staff and patient-patient contacts by design. Assortativity by degree is negative in the total network (-0.13; sd: 0.10), indicating that highly connected individuals are more likely to be in contact with less connected individuals. It is also strongly negative in the patient-staff subgraph (-0.42; sd: 0.14), reflecting the expected disassortivity of healthcare contacts, where each staff member is in contact with multiple patients, whilst each patient is contact with relatively few staff members. In the patient-patient and staff-staff subgraphs, assortativity by degree is positive, as frequently seen in social networks.

Table 1: Summary of network characteristics for the observed total network, patient-patient subgraph, staff-staff subgraph, and patient-staff subgraph. Values were estimated for each day of the 28-days period and summarised here with the mean and standard deviation (sd). Transitivity is not calculated for the patient-staff subgraph as triangles of contacts cannot occur in this network.

	Total	**Patient-patient**	**Staff-staff**	**Patient-staff**
Degree (sd)	12.99 (3.53)	5.25 (1.87)	5.82 (1.87)	8.09 (1.89)
Global efficiency (sd)	0.40 (0.05)	0.25 (0.08)	0.32 (0.10)	0.31 (0.05)
Density (sd)	0.07 (0.01)	0.05 (0.01)	0.09 (0.01)	0.05 (0.00)
Transitivity (sd)	0.37 (0.02)	0.41 (0.05)	0.56 (0.07)	NA
Assortativity (sd)				
By degree	-0.13 (0.10)	0.22 (0.10)	0.14 (0.14)	-0.42 (0.14)
By ward	0.59 (0.08)	0.77 (0.11)	0.72 (0.09)	0.47 (0.09)
Temporal correlation	0.47 (0.11)	0.65 (0.07)	0.35 (0.16)	0.41 (0.12)

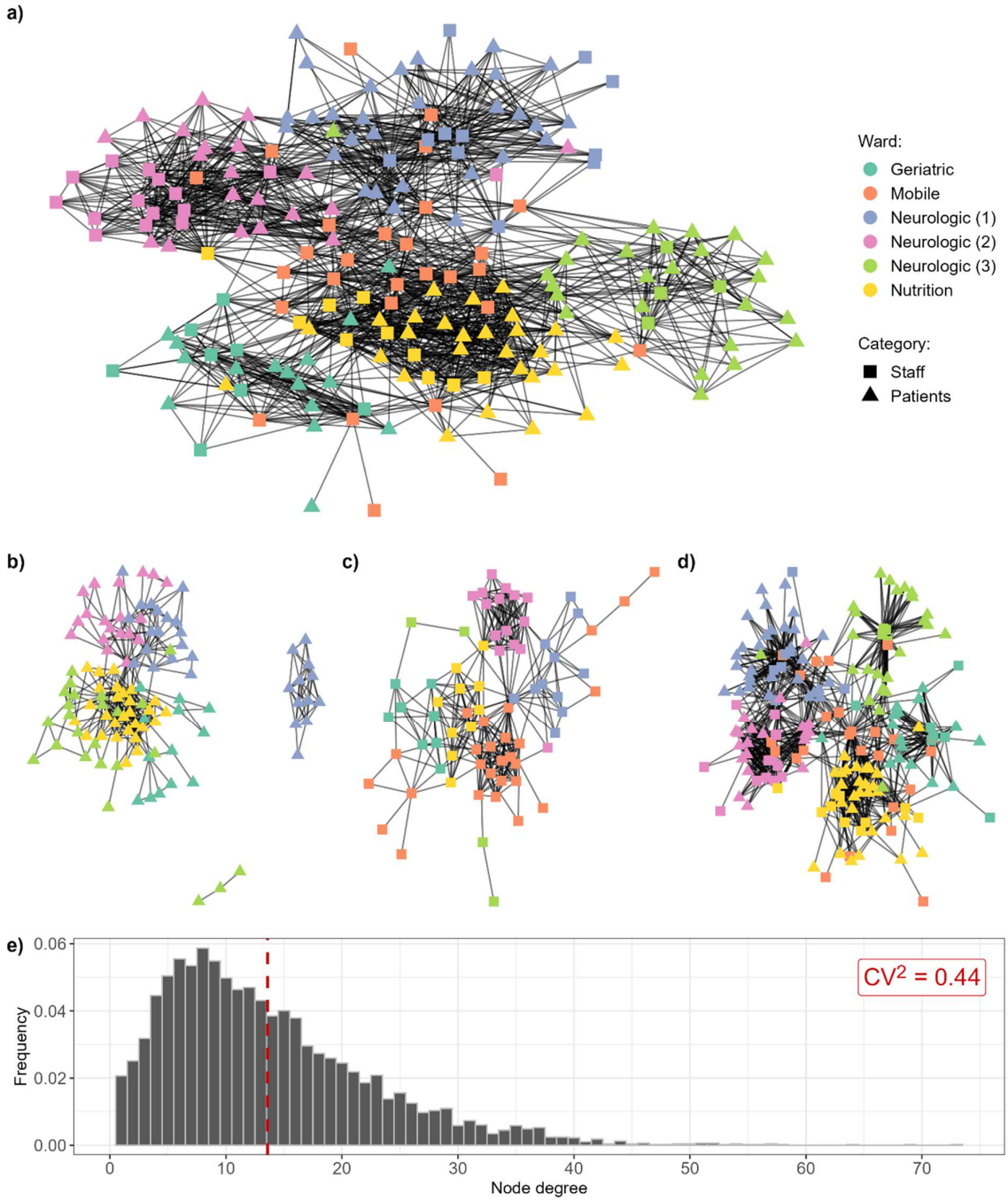
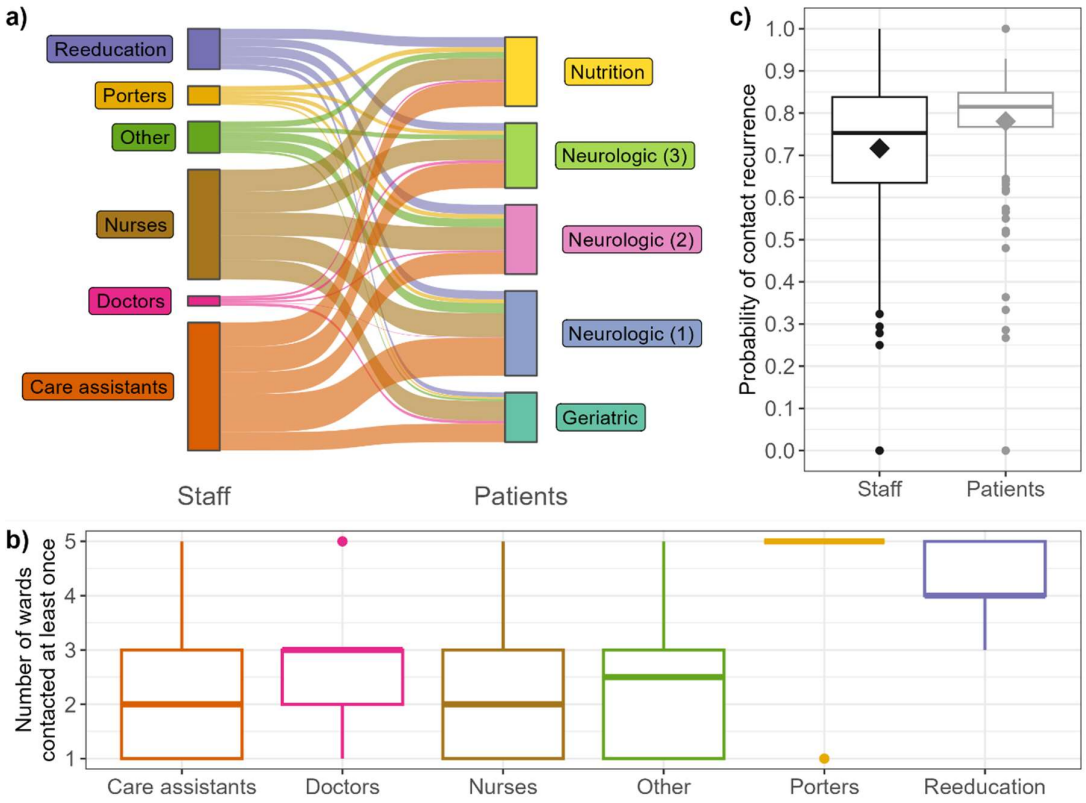


Figure 1: Representation of the observed (a) total network, and (b) patient-patient, (c) staff-staff and (d) patient-staff subgraphs on a single day. The date of 28th of July 2009 was chosen arbitrarily. The layout was calculated using the Kamada-Kawai algorithm, with no weights applied to edges. **e) Distribution of individual degrees for the total network per person per day, across the entire study period.** The dashed red line indicates the mean degree (13.59). CV: coefficient of variation (standard deviation/mean).

Visually, we observe that contacts are naturally clustered by ward (Figure 1a-d). This is reflected in the assortativity by ward, which is systematically high (> 0.45) and indicates that individuals in a ward are always more likely to have contacts with other individuals in the same ward than with individuals in other wards (Table 1). We also observe that contacts exist between all grouped staff professions and patients in different wards, although the distribution is heterogeneous (Figure 2a-b). For example, the median number of wards with which a care assistant (orange) is in contact with is two, while almost all porters (yellow) have contacts with patients from all five wards (Figure 2b).

Overall, contacts are relatively well maintained over time, as shown by the temporal correlation coefficient of 0.47 (sd: 0.11, Table 1). This corresponds to the average probability that, between two subsequent days, an individual maintains the same number of unique contacts, with the same individuals. This metric is highest in the patient-patient subgraph (0.65, sd: 0.07) and lowest in the patient-staff subgraph (0.35, sd: 0.16), indicating that patients tend to have the same contacts with each other every day, whilst contacts amongst healthcare workers often vary between subsequent days. This consistency over time is reflected in the probability of recurring contacts (mean probability: 0.78 for patients, 0.71 for staff), although we note more variability amongst staff than patients (Figure 2c).

All the characteristics described above differ between weekdays and weekends in the network and indicate that there are fewer contacts during weekends (Supplementary Table 1). This difference is reflected in the temporal correlation, which tends to be high when comparing Sunday to Saturday, but low when comparing Saturday to Friday and Monday to Sunday, indicating that the structure of the network changes the most between these timepoints (Supplementary Figure 3).



312

313 **Figure 2: Description of contact heterogeneity and recurrence across the facility. a)**
314 **Repartition of contacts between grouped staff professions and patient wards.** A link
315 between one staff category and one patient ward indicates that, at any point during the
316 investigation period, a staff member from that category had a contact with a patient from that
317 ward. For ease of visualisation, occupational therapists, physiotherapists, and other re-
318 education staff are grouped into “Reeducation”; administrative, animation/hairdresser,
319 logistic, and hospital service agents are grouped into “Other”; and nurses, head nurses, and
320 students/interns are grouped into “Nurses”. Porters, doctors and care assistants are not
321 grouped. **b) Distribution of number of wards with which each staff member has had at least**
322 **one contact with during the study period. c) Distribution of probabilities of recurring**
323 **contacts.** Each observation is calculated over the entire period, and corresponds to the
324 average probability for one staff or one patient to form a new contact with a previously-met
325 individual (staff or patient) rather than a new individual. Diamonds indicate the mean values.

326

327

328

Comparison of synthetic networks

We applied our algorithm to the network described above to stochastically construct four types of synthetic networks: 100 full reconstructed networks, 100 reconstructed networks with observation bias, 100 full pseudo-random networks, and 100 pseudo-random networks incorporating observation bias. We expected that the characteristics of the reconstructed networks with observation bias would be broadly similar to those of the observed i-Bird network. Summary network characteristics are reported in Figure 3 and Supplementary Figure 4.

The daily degrees in the reconstructed networks were slightly higher than the observed network (Figure 3a). Global efficiency was similar between the observed and reconstructed networks, but slightly higher in the reconstructed network with bias (Figure 3b). This is because the algorithm with bias removed individuals from the network at times when they did not wear their sensor during the study, hence reducing the average distance between remaining individuals. For the same reason, the density of the reconstructed network with bias was slightly higher than the observed (Supplementary Figure 4). Transitivity was slightly higher for the reconstructed network with observation bias than without, but lower than the observed network in any case (Figure 3c), as expected since the algorithm did not take into account any element of transitivity when constructing synthetic networks. Finally, assortativity by degree and by ward, as well as temporal correlation, were all well preserved in the reconstructed networks (Figure 3d-f). As a comparison, the random networks with or without bias either substantially over- or under-estimated the values for all metrics compared to the observed network (Figure 3a-f), although we note that transitivity was similar to the other synthetic networks (Figure 3c).

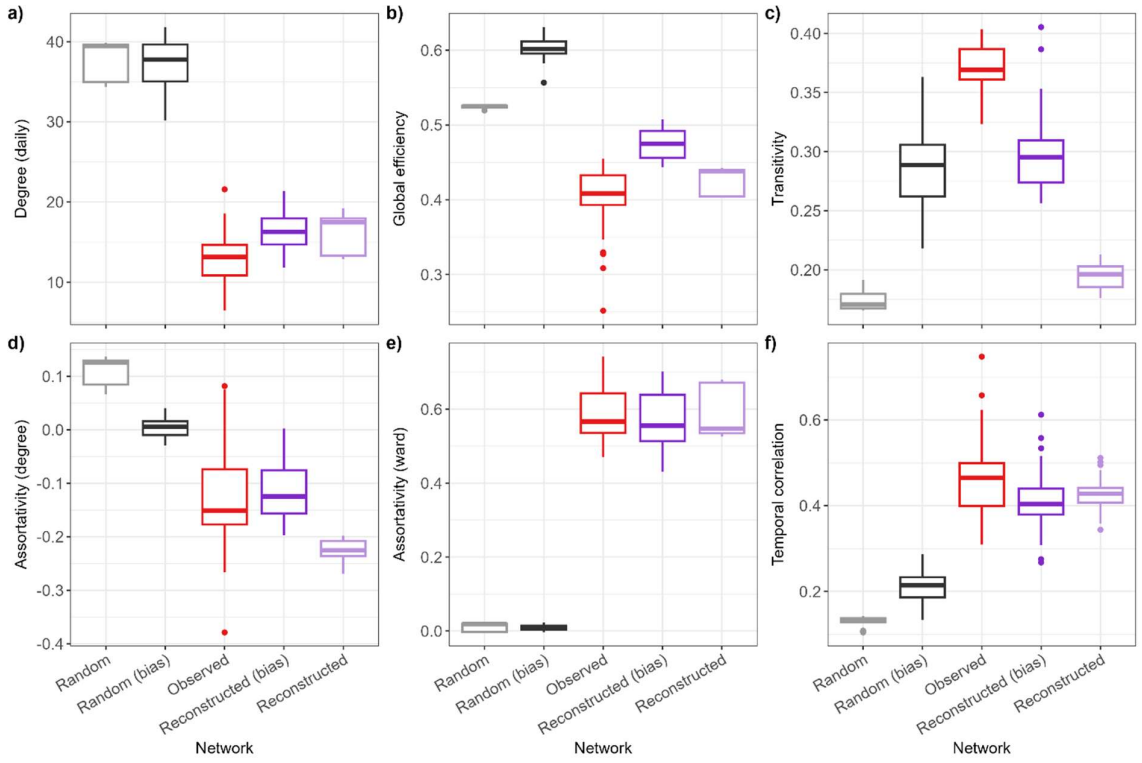
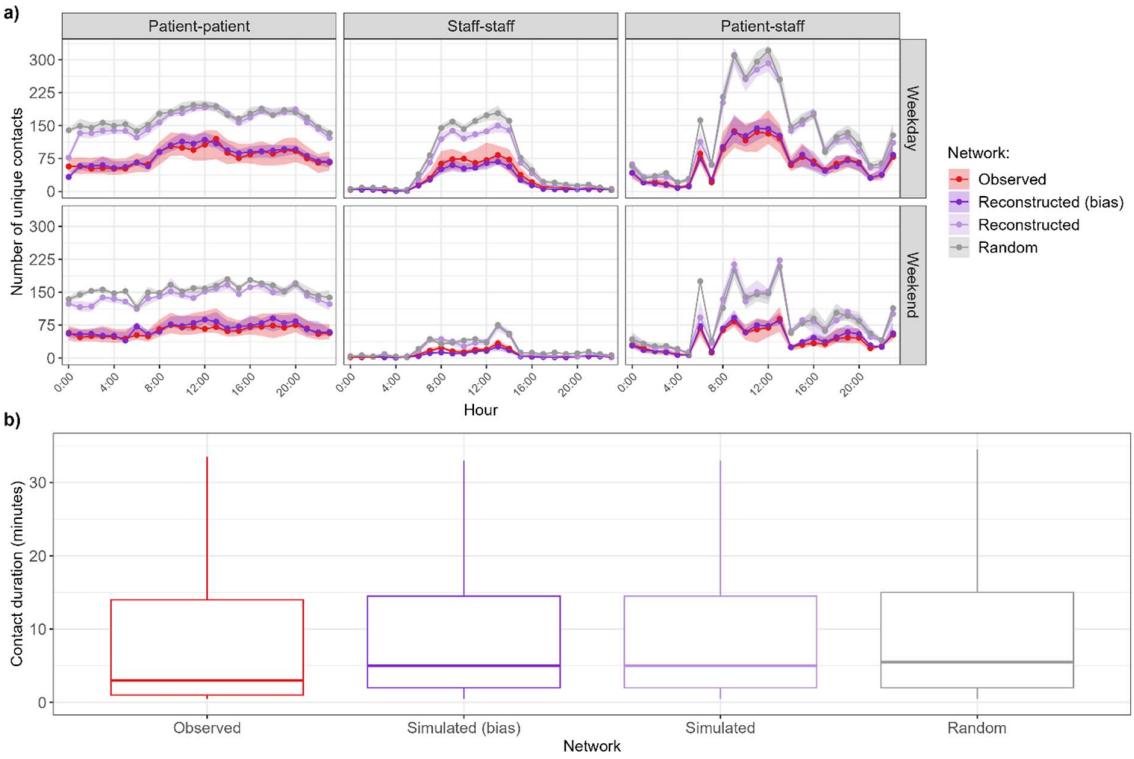


Figure 3: Comparison of network characteristics. The reconstructed networks with observation bias exclude individuals from the network at times when they were known to not wear their sensors. The random networks did not take into account the ward-level structure of the contacts or the probability of recurring contacts. Boxplots for the observed network show the distribution of values calculated for each day. Boxplots for all reconstructed and random networks show the distribution of the median values calculated for each day across 100 networks.

The hourly distributions of numbers of unique patient-patient, staff-patient and staff-staff contacts in the reconstructed network with bias align with those in the observed network (Figure 4a). Whilst these two networks are only partially observed since individuals in the i-Bird study did not have contacts recorded during all their presence days, those unobserved contacts are present in the reconstructed network without bias, leading to approximately twice as many contacts in that network (Figure 4a). Similarly, the random network without bias which is only informed by the hourly distribution of patient-patient, staff-staff and patient-staff contact rates is aligned with the reconstructed network (Figure 4a).

373 Finally, the distributions of contact durations in the synthetic networks were similar to the
374 distribution in the observed network, although there are slightly less shorter contacts (Figure
375 4b). This is because all networks sample their contact durations from a lognormal distribution
376 parameterised by the mean and variance estimated from the data, which puts less emphasis
377 on very short contacts of less than one minute (Supplementary Figure 5).
378



379

380 **Figure 4: Comparison of network contact number and duration. a) Distribution of number**
381 **of unique contacts per hour, separated by type of day (weekday or weekend).** Points
382 correspond to the median, and the shaded areas correspond to the interquartile range. **b)**
383 **Distribution of contact durations.** For ease of visualisation, outliers are not shown on the
384 graph.
385

386 Finally, in supplementary analyses, we assessed the robustness of our algorithm by
387 quantifying the variability of network characteristics across 100 reconstructed networks
388 without bias (Supplementary Figure 6). The variability across reconstructed networks was not
389 statistically significant for any metric (Kruskal-Wallis test, p value > 0.05) except for
390 assortativity by degree ($p < 0.001$). We also aimed to validate our approach by generating "re-
391 simulated" networks informed by summary statistics derived from the reconstructed

392 networks with bias. These re-simulated networks are similar to the full reconstructed
393 networks, indicating that our algorithm consistently recreates realistic networks and
394 reconstructs unobserved contacts (Supplementary Figure 7). However, the number of patient-
395 patient contacts in the re-simulated networks is slightly higher than in the reconstructed
396 networks (Supplementary Figure 7).

Discussion

Summary of findings

In this article, we present an approach to construct stochastic synthetic temporal contact networks from partially observed contact data. To test our algorithm, we used 28 days of CPI data from the i-Bird study, which recorded contacts of staff and patients in a long-term care facility. The observed i-Bird network was heterogeneous, with notably a strong assortativity by ward, varying contact rates between different staff categories and patients, and different contact structures on weekends compared to weekdays. Importantly, we observed temporal correlation between subsequent days in the network, and we estimated that individuals were generally more likely to have contacts with other individuals they previously met rather than new individuals. We therefore informed our reconstruction algorithm with both contact rates by hour, type of day (weekday or weekend) and staff category, and probabilities of recurring contacts estimated for patients and staff using the i-Bird data. The resulting reconstructed networks reproduced well the characteristics of the observed network, as well as the specific distribution of unique contacts per hour.

Similarities between the observed and reconstructed networks

The i-Bird contact network was only partially observed, since contact data was missing on average for 37% of each patient's presence days (42% for staff). This could have occurred for a number of reasons which we cannot distinguish, including depleted batteries, sensor malfunction, or imperfect sensor-wearing compliance. However, since the average contact rates of individuals did not correlate with the proportion of their presence time during which no contact data were recorded (Supplementary Figure 2), it can be assumed that contact patterns on unobserved days were similar to those on observed days. With that assumption, we were able to reconstruct contacts at those times when individuals were present but had no reported contact data. The resulting full reconstructed network is a valuable representation of individual interactions, as it represents the "true" contact network, compared to the i-Bird empirical network which was only partially observed. Although we were inherently limited in our ability to validate this network since the real, fully observed network was not available, we compared it to a re-simulated network which used the

reconstructed network with observation bias as input. The reconstructed and re-simulated networks without bias were almost identical with regards to all the network metrics we considered (Supplementary Figure 7), demonstrating the consistency of our algorithm to reconstruct contacts.

To generate a reconstructed network directly comparable to the observed i-Bird data, we included an observation process to only simulate contacts for individuals at hours when they had contact data reported. This reconstructed network with bias and the observed network had similar positive assortativity by ward, as expected since the input data captured the contact structure by ward. The negative assortativity by degree was also similar, however we noted variability between different networks generated independently by the algorithm (Supplementary Figure 6). Since the algorithm did not directly account for assortativity when simulating networks, this similarity stems from our use of a recurring contact probability coupled with the contact rates estimated by staff categories, resulting in a non-random contact structure with regards to this metric. The hourly contact distribution of patient-patient, staff-staff, and patient-staff contacts was also successfully reproduced by our algorithm.

A key metric of interest here is temporal correlation, which indicates how conserved the network structure is over time. This type of metric is useful to determine the efficiency of disease spread across temporal networks over time [29,30]. Since our algorithm took into consideration the probability of recurring contacts between individuals, our reconstructed networks displayed similar temporal correlation as observed, whilst random networks substantially underestimated this. This aspect is therefore an important strength of our approach, compared to only using estimated average contact rates to construct synthetic contacts.

Limitations of the algorithm

Density and global efficiency in the reconstructed network with bias were slightly higher than in the observed network. This is a likely consequence of our observation process which forcibly removed individuals from the network at times when they had no contacts recorded, hence

458 reducing the number of nodes available in the network. Simultaneously, there was still a need
459 at those times to generate some novel contacts between individuals who never previously
460 met, since the probability of recurring contacts was less than 1. Combined, these elements
461 increased the overall connectivity amongst all individuals in the reconstructed network with
462 bias. Although this could facilitate disease transmission across these reconstructed networks
463 if they are used for such purpose [31], the high assortativity by ward may counter this effect
464 by slowing down transmission across the entire healthcare facility.
465
466 Our algorithm did not specifically account for transitivity when recreating contacts. This is
467 likely why the resulting transitivity was similar to that of the random network and
468 underestimated the observed value (Figure 3). Similarly to density and global efficiency
469 mentioned above, any transitivity in the reconstructed network was likely an indirect
470 consequence of assortativity by ward, restricting the pool of available individuals to generate
471 contacts and leading to interconnectivity between individuals present in the same ward. In
472 any case, this may not substantially affect disease transmission simulated across these
473 networks, since previous work has shown that transitivity is a poor predictor of the total
474 number of individuals who would be infected across the network [31]. Whilst we could extend
475 our algorithm to consider transitivity when choosing the individuals to put in contact, we
476 decided not to do this here to maximise the generalisability of our approach by not requiring
477 such highly detailed contact data.
478
479 Information about patient temporary releases from the hospital (e.g., for weekends with their
480 families) was not available in the i-Bird data, hence such events were not accounted for here,
481 although they may occur frequently in a LTCF. Consequently, the duration of hospital stays
482 may have been overestimated, leading to an overestimation of some contacts among
483 patients. Although this is negligible when comparing the observed and reconstructed network
484 with bias, this is likely why the re-simulated networks slightly overestimated the number of
485 patient-patient contacts compared to the full reconstructed network (Supplementary Figure
486 7). We expect that this overestimation would be absent in settings with more complete
487 information on individual presence, such as schools, workplaces, or acute care facilities.
488
489 **Future work**

In this study, we restricted our detailed analysis of the accuracy of our algorithm to a long-term care facility setting. However, contact structures are known to vary depending on the setting investigated, even between different healthcare settings such as long-term or acute. In addition, even though we had substantial data covering four weeks, it's unclear how representative this contact structure is for other time periods. Notably, our study period falls in the middle of the summer, which is a holiday period in France and may have affected contact patterns. A first important next step would therefore be to repeat this analysis using data collected in a different setting such as acute care, over a different time period. Although we do not expect that our algorithm will perform differently, the strengths and limitations we have highlighted above may be more or less relevant in these different settings. For example, in a setting with low transitivity, the fact that our algorithm underestimates this metric would be less problematic.

Here we directly re-used patient admission and discharge data as well as staff schedules to identify which individuals were present in the facility at each hour, and hence whom the algorithm had to build contacts for. While this choice was coherent since our aim was to compare the observed and reconstructed networks, a second possible extension of our work would be to simulate the presence of individuals over time. This could be implemented by extracting admission and discharge rates for each category of staff and patients and using these values to recreate new presence times for individuals by sampling from relevant probability distributions. This would allow us to further account for possible variability in the structure of the population in the facility, and hence add further stochasticity in our algorithm.

Since contact data may only be available for short periods of time, a third question of interest would be to understand the volume of data required to generate realistic temporal contact networks using our algorithm. In our main analysis, we used the entire four weeks available to both derive contact parameters and compare the reconstructed and observed networks. For sensitivity, we also considered smaller timer periods to calculate the summary contact parameters required by the algorithm (Supplementary Figure 8). As expected, this led to variability amongst the reconstructed networks depending on the length of the period used. In any case, the main risk of using only a short period of time is to miss out some contacts between categories. For example, during a single week, by chance there may not be any

observed contact between patients from one ward *w1* and a nurse from another ward *w2*, while in reality over a longer period of time we may observe a few of such contacts. In that case, the algorithm will systematically assume that such contacts never occur during the entire period over which the reconstructed networks are generated and will therefore construct an incomplete network. Therefore, it is essential for users to be confident that the data they use include contact rates for all relevant categories in their setting and for typical representative days.

As discussed above, taking into account the probability for contacts to be recurring instead of assuming a uniform distribution is a key element of our approach. Here, we estimate the average probabilities of recurring contacts as 0.71 for staff and 0.78 for patients, but we note some individual variation in this value (Figure 2). In addition, our estimation here is made using the entire observed contact networks, but this may be difficult in instances where only limited data are available. For sensitivity, we investigated the impact of manually setting the probabilities to 0.1, 0.5 and 0.9 for both staff and patients (Supplementary Figure 9). This led to important variations in assortativity by degree and temporal correlation compared to using the estimated probability. A greater understanding of this recurring contact probability in various settings would be helpful to improve the generalisability of our algorithm and will also be useful more broadly to better understand contact heterogeneity. In healthcare settings, this probability could likely be estimated without requiring complete contact data, using information on staff schedules and patient ward allocation instead.

Deep learning algorithms such as graph convolutional networks (GCN) have become increasingly popular to study contact networks, particularly in the context of infectious disease transmission [32–36]. It would be interesting to compare the performance of these approaches with our algorithm to estimate network characteristics and reconstruct unobserved contacts. However, traditional GCN approaches do not account for temporal dependencies between contacts such as the ones we observed in the i-Bird network where the probability of recurring contacts plays a key role [37,38]. On the other hand, temporal graph networks can capture this temporal dependency [39,40], but require substantial computational resources to be applied to a network such as i-Bird, with hundreds of interactions recorded every 30 seconds during several weeks. Finally, deep learning methods

554 require large amounts of training data. Democratising their use would therefore first require
555 new studies to collect close-proximity interaction data in different settings and time periods,
556 presenting further logistical challenges.
557
558 ## Implications
559 Our algorithm relied on computing summary statistics from an observed network, then using
560 these statistics to stochastically reconstruct contact networks. Such statistics could be derived
561 directly from other observed networks, as we have done here to validate our approach. In
562 that case, our approach to generate multiple reconstructed networks could be useful to
563 evaluate the impact of stochasticity on the contact structure in a given setting, instead of only
564 relying on a single observed network. Our approach could also be used to infer information
565 on potentially unobserved contacts and to predict realistic temporal dynamics over longer
566 time periods than the data collection.
567
568 Alternatively, summary contact statistics could be more simply collected from cross-sectional
569 surveys or even derived exclusively from individual schedules, which would not require a
570 detailed and costly follow-up using sensors. In this scenario, the only other data required
571 would be individual presence times, which should either be routinely available (e.g. in
572 healthcare settings or schools) or relatively easy to collect (e.g. in workplaces). Although as
573 mentioned in the Limitations the amount of data our algorithm requires to generate realistic
574 networks is still unclear, our approach could ultimately be used to generate contact networks
575 from contact matrices. This would substantially facilitate research on the impact of contact
576 heterogeneity in various populations and settings.
577
578 To the best of our knowledge, this is the first proposed algorithm to generate temporal contact
579 networks by taking into consideration empirically measured contact rates based on close-
580 proximity sensors, while most available packages only construct static networks and rely on
581 hyperparameters [41]. These temporal networks could then be used within mathematical
582 models used to evaluate the potential impact of interventions against disease transmission
583 networks [9–12]. In particular, this could improve the wider applicability of individual-based

584 model which can take into account this detailed contact heterogeneity to test the effect of
585 highly targeted interventions.

References

1. Keeling MJ, Rohani P. Modeling Infectious Diseases in Humans and Animals. Modeling Infectious Diseases in Humans and Animals. Princeton University Press; 2008. doi:10.1515/9781400841035

2. Anderson RM, May RM. Infectious diseases of humans: dynamics and control. Reprinted. Oxford: Oxford Univ. Press; 2010.

3. Diekmann O, Heesterbeek JAP. Mathematical Epidemiology of Infectious Diseases: Model Building, Analysis and Interpretation. John Wiley & Sons; 2000.

4. Großmann G, Backenköhler M, Wolf V. Heterogeneity matters: Contact structure and individual variation shape epidemic dynamics. PLoS One. 2021;16: e0250050. doi:10.1371/journal.pone.0250050

5. Kiss IZ, Miller JC, Simon PL, others. Mathematics of epidemics on networks. Cham: Springer. 2017;598: 31.

6. Danon L, Ford AP, House T, Jewell CP, Keeling MJ, Roberts GO, et al. Networks and the Epidemiology of Infectious Disease. Interdiscip Perspect Infect Dis. 2011;2011: 284909. doi:10.1155/2011/284909

7. Mossong J, Hens N, Jit M, Beutels P, Auranen K, Mikolajczyk R, et al. Social Contacts and Mixing Patterns Relevant to the Spread of Infectious Diseases. PLOS Medicine. 2008;5: e74. doi:10.1371/journal.pmed.0050074

8. Keeling MJ, Eames KTD. Networks and epidemic models. Journal of The Royal Society Interface. 2005;2: 295–307. doi:10.1098/rsif.2005.0051

9. Bansal S, Read J, Pourbohloul B, Meyers LA. The dynamic nature of contact networks in infectious disease epidemiology. Journal of Biological Dynamics. 2010;4: 478–489. doi:10.1080/17513758.2010.503376

10. Masuda N, Holme P. Predicting and controlling infectious disease epidemics using temporal networks. F1000Prime Rep. 2013;5: 6. doi:10.12703/P5-6

11. Gross T, D'Lima CJD, Blasius B. Epidemic Dynamics on an Adaptive Network. Phys Rev Lett. 2006;96: 208701. doi:10.1103/PhysRevLett.96.208701

12. Valdano E, Poletto C, Giovannini A, Palma D, Savini L, Colizza V. Predicting Epidemic Risk from Past Temporal Contact Data. PLOS Computational Biology. 2015;11: e1004152. doi:10.1371/journal.pcbi.1004152

13. Holme P, Saramäki J. Temporal networks. Physics Reports. 2012;519: 97–125. doi:10.1016/j.physrep.2012.03.001

14. Hornbeck T, Naylor D, Segre AM, Thomas G, Herman T, Polgreen PM. Using Sensor Networks to Study the Effect of Peripatetic Healthcare Workers on the Spread of

Hospital-Associated Infections. Journal of Infectious Diseases. 2012;206: 1549–1557. doi:10.1093/infdis/jis542

15. Obadia T, Silhol R, Opatowski L, Temime L, Legrand J, Thiébaut ACM, et al. Detailed Contact Data and the Dissemination of Staphylococcus aureus in Hospitals. Salathé M, editor. PLoS Comput Biol. 2015;11: e1004170. doi:10.1371/journal.pcbi.1004170

16. Min-Allah N, Alahmed BA, Albreek EM, Alghamdi LS, Alawad DA, Alharbi AS, et al. A survey of COVID-19 contact-tracing apps. Computers in Biology and Medicine. 2021;137: 104787. doi:10.1016/j.compbiomed.2021.104787

17. Eames K, Bansal S, Frost S, Riley S. Six challenges in measuring contact networks for use in modelling. Epidemics. 2015;10: 72–77. doi:10.1016/j.epidem.2014.08.006

18. Read JM, Edmunds WJ, Riley S, Lessler J, Cummings D a. T. Close encounters of the infectious kind: methods to measure social mixing behaviour. Epidemiology & Infection. 2012;140: 2117–2130. doi:10.1017/S0950268812000842

19. Smieszek T, Castell S, Barrat A, Cattuto C, White PJ, Krause G. Contact diaries versus wearable proximity sensors in measuring contact patterns at a conference: method comparison and participants' attitudes. BMC Infectious Diseases. 2016;16: 341. doi:10.1186/s12879-016-1676-y

20. Watts DJ, Strogatz SH. Collective dynamics of 'small-world' networks. Nature. 1998;393: 440–442. doi:10.1038/30918

21. Almutiry W, Deardon R. Contact network uncertainty in individual level models of infectious disease transmission. Stat Commun Infect Dis. 2021;13: 20190012. doi:10.1515/scid-2019-0012

22. Shirley MDF, Rushton SP. The impacts of network topology on disease spread. Ecological Complexity. 2005;2: 287–299. doi:10.1016/j.ecocom.2005.04.005

23. Gimma A, Munday JD, Wong KLM, Coletti P, van Zandvoort K, Prem K, et al. Changes in social contacts in England during the COVID-19 pandemic between March 2020 and March 2021 as measured by the CoMix survey: A repeated cross-sectional study. PLoS Med. 2022;19: e1003907. doi:10.1371/journal.pmed.1003907

24. Mousa A, Winskill P, Watson OJ, Ratmann O, Monod M, Ajelli M, et al. Social contact patterns and implications for infectious disease transmission – a systematic review and meta-analysis of contact surveys. Rodriguez-Barraquer I, Serwadda DM, editors. eLife. 2021;10: e70294. doi:10.7554/eLife.70294

25. Hoang T, Coletti P, Melegaro A, Wallinga J, Grijalva CG, Edmunds JW, et al. A Systematic Review of Social Contact Surveys to Inform Transmission Models of Close-contact Infections. Epidemiology. 2019;30: 723–736. doi:10.1097/EDE.0000000000001047

26. Duval A, Obadia T, Martinet L, Boëlle P-Y, Fleury E, Guillemot D, et al. Measuring dynamic social contacts in a rehabilitation hospital: effect of wards, patient and staff characteristics. Scientific Reports. 2018;8: 1686. doi:10.1038/s41598-018-20008-w

27. R Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing; 2022. Available: https://www.R-project.org/

28. Csardi G, Nepusz T. The igraph software package for complex network research. InterJournal. 2006;Complex Systems: 1695.

29. Tang J, Scellato S, Musolesi M, Mascolo C, Latora V. Small-world behavior in time-varying graphs. Phys Rev E. 2010;81: 055101. doi:10.1103/PhysRevE.81.055101

30. Kretzschmar M, Morris M. Measures of concurrency in networks and the spread of infectious disease. Mathematical Biosciences. 1996;133: 165–195. doi:10.1016/0025-5564(95)00093-3

31. Pérez-Ortiz M, Manescu P, Caccioli F, Fernández-Reyes D, Nachev P, Shawe-Taylor J. Network topological determinants of pathogen spread. Sci Rep. 2022;12: 7692. doi:10.1038/s41598-022-11786-5

32. Fritz C, Dorigatti E, Rügamer D. Combining graph neural networks and spatio-temporal disease models to improve the prediction of weekly COVID-19 cases in Germany. Sci Rep. 2022;12: 3930. doi:10.1038/s41598-022-07757-5

33. Gao J, Sharma R, Qian C, Glass LM, Spaeder J, Romberg J, et al. STAN: spatio-temporal attention network for pandemic prediction using real-world evidence. Journal of the American Medical Informatics Association. 2021;28: 733–743. doi:10.1093/jamia/ocaa322

34. Panagopoulos G, Nikolentzos G, Vazirgiannis M. Transfer Graph Neural Networks for Pandemic Forecasting. Proceedings of the AAAI Conference on Artificial Intelligence. 2021;35: 4838–4845. doi:10.1609/aaai.v35i6.16616

35. Kapoor A, Ben X, Liu L, Perozzi B, Barnes M, Blais M, et al. Examining COVID-19 Forecasting using Spatio-Temporal Graph Neural Networks. arXiv; 2020. doi:10.48550/arXiv.2007.03113

36. Zhao G, Jia P, Zhou A, Zhang B. InfGCN: Identifying influential nodes in complex networks with graph convolutional networks. Neurocomputing. 2020;414: 18–26. doi:10.1016/j.neucom.2020.07.028

37. Li L, Zhou J, Jiang Y, Huang B. Propagation source identification of infectious diseases with graph convolutional networks. Journal of Biomedical Informatics. 2021;116: 103720. doi:10.1016/j.jbi.2021.103720

691 38. Ni Q, Wu X, Chen H, Jin R, Wang H. Spatial-temporal deep learning model based rumor
692 source identification in social networks. J Comb Optim. 2023;45: 86.
693 doi:10.1007/s10878-023-01018-5

694 39. Holme P. Modern temporal network theory: a colloquium. Eur Phys J B. 2015;88: 234.
695 doi:10.1140/epjb/e2015-60657-4

696 40. Tang J, Leontiadis I, Scellato S, Nicosia V, Mascolo C, Musolesi M, et al. Applications of
697 Temporal Graph Metrics to Real-World Networks. In: Holme P, Saramäki J, editors.
698 Temporal Networks. Berlin, Heidelberg: Springer; 2013. pp. 135–159. doi:10.1007/978-
699 3-642-36461-7_7

700 41. Prettejohn B, Berryman M, McDonnell M. Methods for Generating Complex Networks
701 with Selected Structural Properties for Simulations: A Review and Tutorial for
702 Neuroscientists. Frontiers in Computational Neuroscience. 2011;5. Available:
703 https://www.frontiersin.org/articles/10.3389/fncom.2011.00011

704