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

Audrey Duval, Quentin Leclerc, Didier Guillemot, Laura Temime, Lulla Opatowski

PLOS COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

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

Audrey Duval[1,2,3☯¤], **Quentin J. Leclerc**[1,2,3☯*], **Didier Guillemot**[1,2,4], **Laura Temime**[3,5‡], **Lulla Opatowski**[1,2‡]

1 Institut Pasteur, Université Paris Cité, Epidemiology and Modelling of Bacterial Escape to Antimicrobials (EMEA), Paris, France, **2** INSERM, Université Paris-Saclay, Université de Versailles St-Quentin-en-Yvelines, Team Echappement aux Anti-infectieux et Pharmacoépidémiologie U1018, CESP, Versailles, France, **3** Laboratoire Modélisation, Epidémiologie et Surveillance des Risques Sanitaires (MESuRS), Conservatoire National des Arts et Métiers, Paris, France, **4** AP-HP, Paris Saclay, Department of Public Health, Medical Information, Clinical research, Garches, France, **5** Institut Pasteur, Conservatoire National des Arts et Métiers, Unité PACRI, Paris, France

☯ These authors contributed equally to this work.
¤ Current address: Imagine Institute, Data Science Platform, INSERM UMR 1163, Université de Paris, Paris, France
‡ LT and LO also contributed equally to this work.
* quentin.leclerc@pasteur.fr

Citation: Duval A, Leclerc QJ, Guillemot D, Temime L, Opatowski L (2024) An algorithm to build synthetic temporal contact networks based on close-proximity interactions data. PLoS Comput Biol 20(6): e1012227. https://doi.org/10.1371/journal.pcbi.1012227

Editor: Eric Lofgren, Washington State University, UNITED STATES

Received: December 8, 2023

Accepted: June 4, 2024

Published: June 13, 2024

Data Availability Statement: The relevant contact networks and analysis code are available in the following GitHub repository: https://github.com/qleclerc/network_algorithm.

Funding: AD, LT and LO received funding from the French National Research Agency (SPHINX-17-CE36-0008-01, https://anr.fr/en/). DG received funding from the National Clinical Research Program and the Investissement d'Avenir program, Laboratoire d'Excellence "Integrative Biology of Emerging Infectious Diseases" (ANR-10-LABX-62-

Abstract

Small populations (e.g., hospitals, schools or workplaces) are characterised by high contact heterogeneity and stochasticity affecting pathogen transmission dynamics. Empirical individual contact data provide unprecedented information to characterize such heterogeneity and are increasingly available, but 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 healthcare 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. Recording bias is formalized through an observation model, to allow direct comparison between the augmented and observed networks. We validate our algorithm using data collected during the i-Bird study, and compare the empirical and reconstructed networks. The algorithm 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 algorithm consistently recreated unobserved contacts to generate full reconstructed networks for the LTCF. To conclude, we propose an approach to generate realistic temporal contact networks and

IBEID). The funders did not play any role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

reconstruct unobserved contacts from summary statistics computed using individual-level interaction networks. This could be applied and extended to generate contact networks to other HCS using limited empirical data, to subsequently inform individual-based epidemic models.

Author summary

Contact networks are the most informative representation of the contact heterogeneity, and therefore infectious disease transmission risk, in small populations. However, the data collection required is costly and complex, usually limited to a few days only and likely to suffer from partially observed data, making the practical integration of networks into models challenging. In this article, we present an approach leveraging empirical individual contact data to stochastically reconstruct realistic temporal networks in healthcare settings. The algorithm accounts for population specificities including the hourly distribution of contact rates between different individuals (staff categories, patients) and the probability for contact repetition between the same individuals. We illustrate and validate this algorithm using a real contact network measured in a long-term care facility. Our approach outperforms random graphs informed by the same data to accurately reproduce observed network characteristics and hourly staff-patient contact patterns. The algorithm recreates unobserved contacts, providing data augmentation for times with missing information. This method should improve the usability and reliability of contact networks, and therefore promote integration of empirical contact data in individual-based models.

Introduction

Limiting the burden of infectious diseases requires a good understanding of how they spread. For pathogens 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 are usually 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 instead of only relying on summary contact rates by groups [5,6].

Contact networks are increasingly used to fully capture the interactions between individuals in small populations [7,8]. These networks explicitly represent the links between all individuals in such populations, as opposed to contact matrices which capture average contact rates between groups of individuals [9,10]. Temporal contact networks further capture the time-changing nature of contacts, therefore representing individual interactions more accurately than static networks [11–15]. 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 [10]. Recently, empirical data collected to build inter-individual temporal networks has become increasingly available to inform contact networks. For example, studies have used sensors to record close-proximity interactions between

individuals [16–18], and contact tracing programs have relied on the integrated Bluetooth technology in mobile phones [19].

However, the detailed empirical data required to build temporal contact networks remain subject to several limitations [20,21]. These data are typically collected over a few days only [22,23], and may be subject to observation bias; sensors might not be properly placed to register contacts [24], or individuals may disable Bluetooth on their mobile phones at different times [19]. 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 [25–27]. 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 [28–30].

Here, we propose an algorithm to stochastically reconstruct realistic contact networks from partially observed contact data in healthcare settings (HCS). To validate our approach, we use close-proximity data collected in a long-term care facility (LTCF) during the i-Bird study [17,31]. We first illustrate the typical complexity of contact structures in HCS through the i-Bird network example. We then compute summary contact parameters from these data to generate reconstructed contact networks and compare these synthetic contact networks with the observed data.

Methods

Building synthetic contacts in a HCS

Algorithm outline. We built an algorithm to stochastically reconstruct a realistic full temporal network of inter-individual close-proximity interactions (CPIs, at less than 1.5m) in a HCS using parameters estimated from empirical individual contact data. This algorithm generates a new synthetic network which notably reconstructs contacts at times when individuals were known to be present in the HCS 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 empirical 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 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:

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

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 ward W_1 and individuals from category C_2 belonging to 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 individual i. 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 between them on instance k of the hour h: it equals 1 if i and j had at least one contact recorded at that time, 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 l of the hour h, $N_{C1W1,l}$ is the number of individuals from C_1W_1 that had any contact recorded during that hour.

This estimation is 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

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

Where $U_{i,d}$ is the set of unique individuals with whom i had a contact on day d, $U_{i,[d0,d\ [}$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 (d non-included), and the notation $|x|$ indicates the cardinality of the set x. 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 daily probability of recurring contacts for individual i across all days as

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

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

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

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

Generation of synthetic CPIs: number and identity of individuals in contacts. For each hour of our period of interest, we estimate the number of contacts between individuals present in the HCS during that hour, determined using admission data and staff schedule. 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' new individuals from those available, in order to avoid double counting.

These 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 Z_i, the subset of C_2S_2 individuals who previously met i, according to probability $p_{i \to j}$:

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

Where $N_{i\to j}$ is the number of previous contacts between i and j before hour h, and $\sum_{k\in S} N_{i\to k}$ is the number of previous contacts between i and each individual k belonging to Z_i.

- Otherwise, the contact is not recurring: the individual j in contact is randomly and uniformly chosen among Z_i', the subset of C_2S_2 individuals who have not yet met i.

Generation of contact durations. For each contact between two given individuals i from C_1S_1 and j from C_2S_2 the duration of contact is sampled from a log-normal distribution calibrated from the observed mean and variance of contact durations between individuals from C_1S_1 and C_2S_2 on hour h.

Validation dataset: the i-Bird network

Dataset description. We validate our algorithm by applying it to data collected during the Individual-Based Investigation of Resistance Dissemination (i-Bird) study [17,31]. 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 CPIs 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 LTCF 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 considered here that mobile staff belonged to an "artificial" 6th ward, to compute contact rates according to the algorithm detailed above. Staff were divided into 13 professions: administrative, animation/hairdresser, logistic, hospital service agent, porter, occupational therapist, physiotherapist, other rehabilitation staff, nurse, head nurse, care assistant, medical student/resident, 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 and compared these schedules to the dates and times when staff had any contact recorded. We assumed that, in reality, staff would have at least one contact with any other individual during any given hour of their presence time, hence if no contact was recorded for a given hour of presence we considered this was missing data rather than true absence of contact. Through this, we estimated that the median percentage of a staff's total presence time when no contact data was recorded was 40.0% (interquartile range (IQR): 0–75.0%). We repeated this analysis for patients at the daily instead of hourly level, as we only had access to admission and discharge dates for patients. We estimated that the median time when no contact data was recorded was 33.3% (interquartile range (IQR): 10.5–53.6%) of a patient's presence days.

Although the overall compliance was high (90% of individuals agreed to wear a sensor), there was therefore substantial heterogeneity in the individual coverage of the raw i-Bird network (S1 Fig). Interestingly, 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, nor their total presence time (S2 Fig).

Observation bias process. As mentioned earlier, the observed i-Bird network, as any real-life data, includes recording biases leading to some periods of non-recording of CPIs, with the extent of this bias varying between individuals. To make our reconstructed networks comparable to the observed one, we therefore introduced an observation bias process. For each individual in the observed network, we identified the hours with no contact recorded. We then removed those individuals on those hours before proceeding with the algorithm described above. The resulting "reconstructed biased network" and the observed network hence suffer from the same bias and are comparable.

Simulations and analysis

From the analysis of the i-Bird empirical network and data, we used our algorithm to generate 100 full synthetic reconstructed networks, and 100 reconstructed networks with observation bias. For comparison, we also used our algorithm to generate 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. The patient-patient, staff-staff, and patient-staff contact rates are calculated as detailed in the section "Estimation of contact rates from the data", treating all staff as if they were part of the same profession, and all individuals as if they were part of the same ward. At each contact, the individual encountered is therefore chosen randomly from all those present in the LTCF at that time, regardless of whether or not the individual was previously encountered. This approach to generate pseudo-random networks is similar to a stochastic block model with two sub-populations (patients and staff).

We implemented the algorithm in C++ with the repast HPC 2.3.0 library. All network reconstructions were performed on the Maestro cluster hosted by the Institut Pasteur. The networks were analysed and the epidemics were simulated in R [32], using the igraph package [33]. 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, considering them as empirical networks. This allowed us to generate 100 new full reconstructed networks from fully known networks, and confirm these "re-simulated networks" were similar to the full reconstructed networks generated from the observed data.

Application: transmission of a pathogen over the network

To investigate the impact of the network structure on the predicted dissemination of a pathogen, we implemented a simple susceptible-infected-recovered epidemic transmission model on the networks. Briefly, we aggregated the contact network at the daily level, without considering contact durations for simplicity. We initiated an epidemic by randomly infecting one individual on the first day, then simulating transmission for each day between individuals in contact. When an infected individual had a contact with a susceptible one, there was a stochastic risk of transmission with a probability of 0.05. Each individual remained infectious with an assumed constant transmissibility risk for an average of 5 days (sampled from a Poisson distribution), after which we assumed they were recovered and could not be reinfected. These parameters equate to an average basic reproduction number of 3.25 in the observed i-Bird network (= 0.05 * 5 * 13), since individuals have on average approximately 13 unique contacts per day (see Results). Using these parameters, we simulated complete epidemics over a one month period: in all networks, the epidemic almost becoming extinct in by the end of the period. We

simulated 100 independent stochastic epidemics for each network, and excluded simulations where less than 10 individuals were infected in total.

Results

Description of HCS contact heterogeneity: the example of the i-Bird dataset

In this section, we illustrate the typical complexity of contact structures in HCS using the i-Bird network. While the algorithm makes use of data at the hourly level, 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 [31].

We first summarise the observed temporal network recorded in the LTCF during the i-Bird study, comparing the total daily network and subgraphs with only patient-patient, staff-staff, or patient-staff contacts (Fig 1A–1D). 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 (Fig 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, distance: 4); staff-staff: 0.32 (sd: 0.10, distance: 3.1); patient-staff: 0.31 (sd: 0.05, distance: 3.2)). 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.

Visually, we observe that contacts are naturally clustered by ward (Fig 1A–1D). 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 (Fig 2A–2B). For example, the median number of wards with which a care

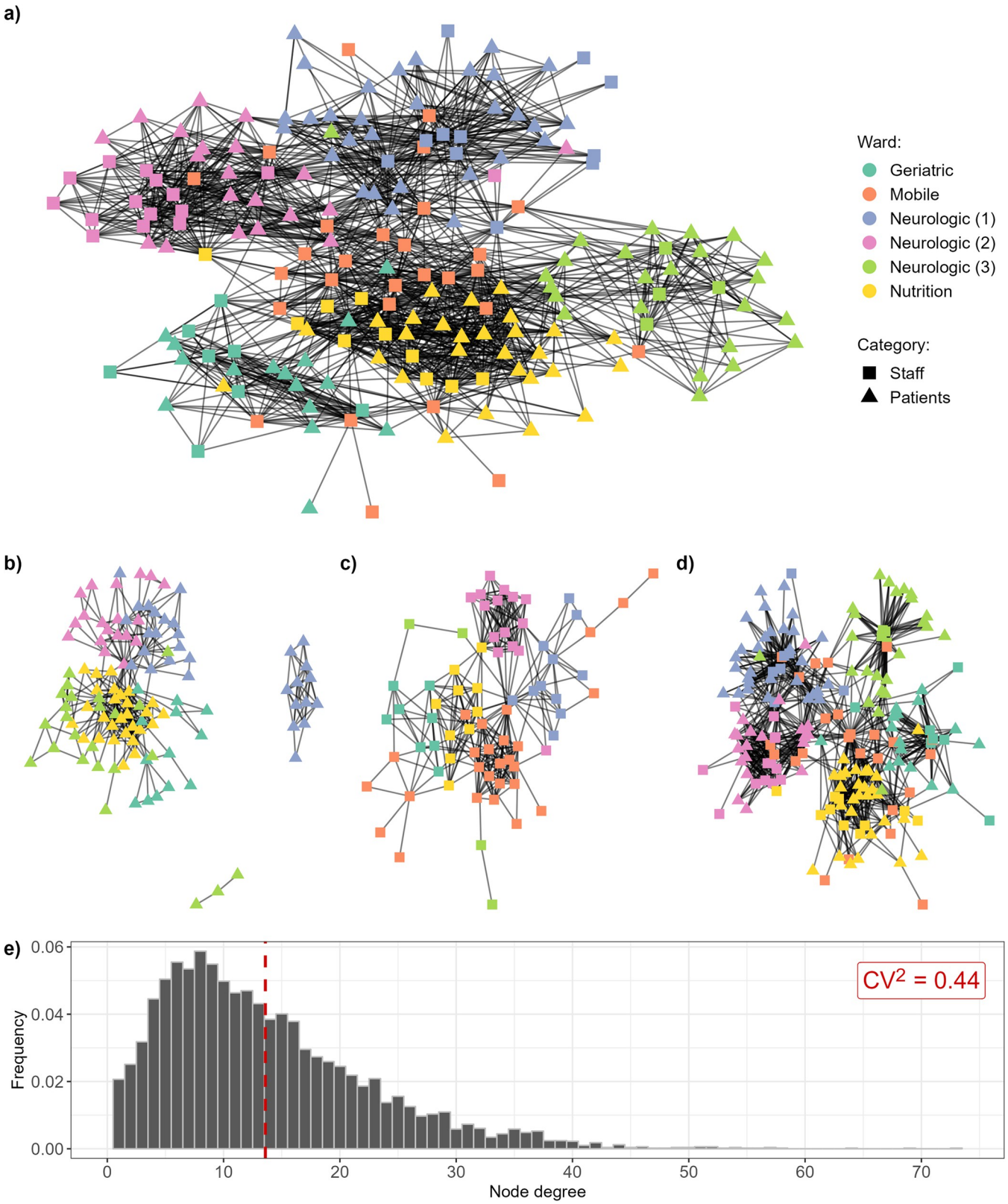


Fig 1. Representation of the observed network recorded during the i-Bird study: (a) total network, and (b) patient-patient, (c) staff-staff and (d) patient-staff subgraphs on a single day. The date of 28th of July 2009 was chosen arbitrarily. The layout was calculated using the Kamada-Kawai algorithm, 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).

https://doi.org/10.1371/journal.pcbi.1012227.g001

assistant (orange) is in contact with is two, while almost all porters (yellow) have contacts with patients from all five wards (Fig 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 high probability of recurring contacts (mean probability: 0.78 for patients, 0.71 for staff), although we note more variability amongst staff than patients (Fig 2C).

All the characteristics described above differ between weekdays and weekends in the network and indicate that there are fewer contacts during weekends (S1 Table). 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 (S3 Fig).

Comparison of synthetic and observed networks

To illustrate and validate our algorithm, we applied it to the i-Bird network described above to stochastically construct four types of synthetic networks using the estimated contact parameters: 100 full reconstructed networks, 100 reconstructed networks incorporating 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 Figs 3 and S4.

The daily degrees in the reconstructed networks were slightly higher than the observed network (Fig 3A). Global efficiency was similar between the observed and reconstructed networks, but slightly higher in the reconstructed network with bias (Fig 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

Table 1. Summary of network characteristics for the observed i-Bird 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)

https://doi.org/10.1371/journal.pcbi.1012227.t001

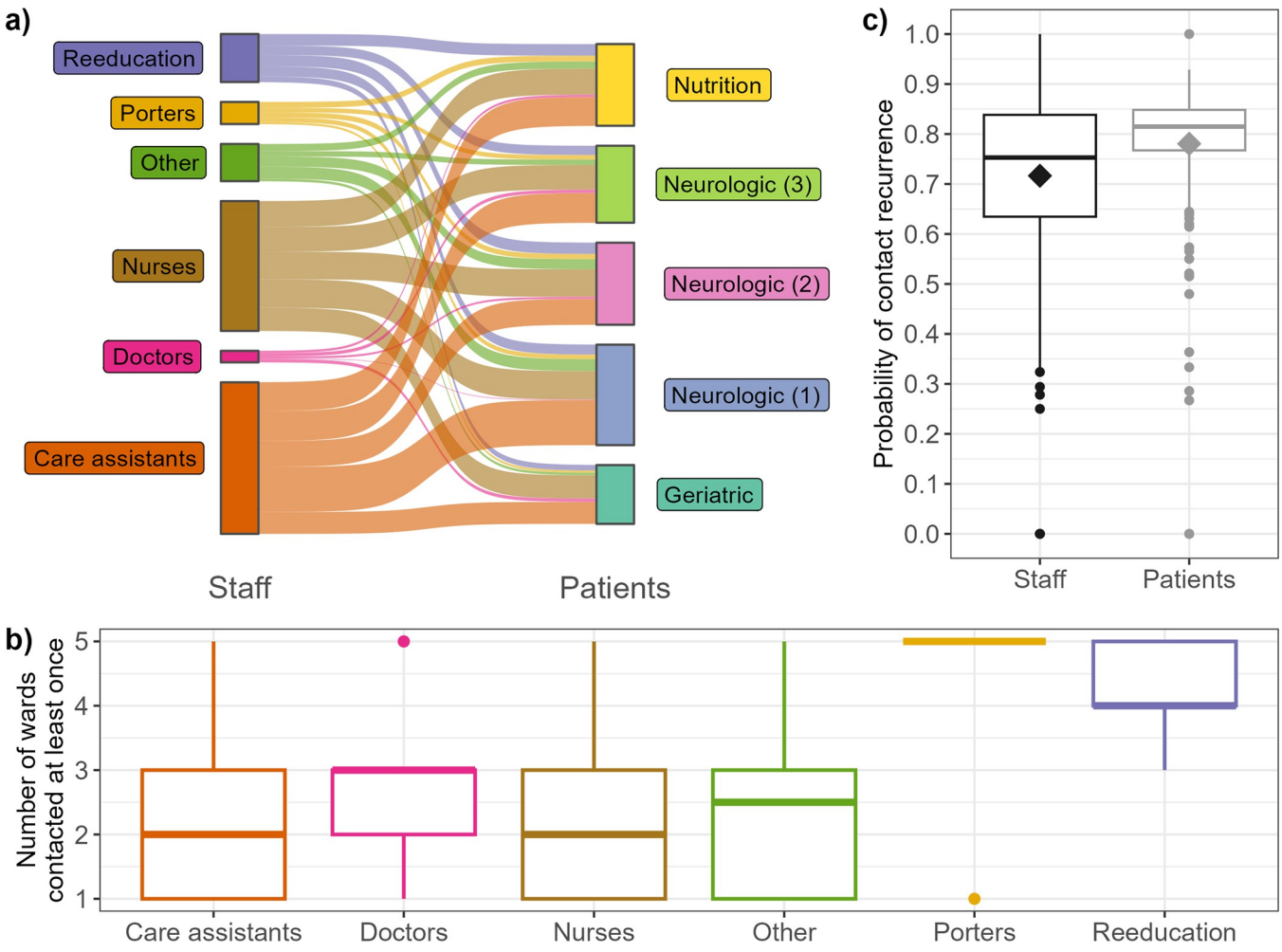


Fig 2. Description of contact heterogeneity and recurrence across the facility. a) Repartition of contacts between grouped staff professions and patient wards. A link between one staff category and one patient ward indicates that, at any point during the investigation period, a staff member from that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

https://doi.org/10.1371/journal.pcbi.1012227.g002

higher than the observed (S4 Fig). Transitivity was slightly higher for the reconstructed network with observation bias than without, but lower than the observed network in any case (Fig 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 (Fig 3D–3F). 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 (Fig 3A–3F), although we note that transitivity was similar to the other synthetic networks (Fig 3C).

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 (Fig 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

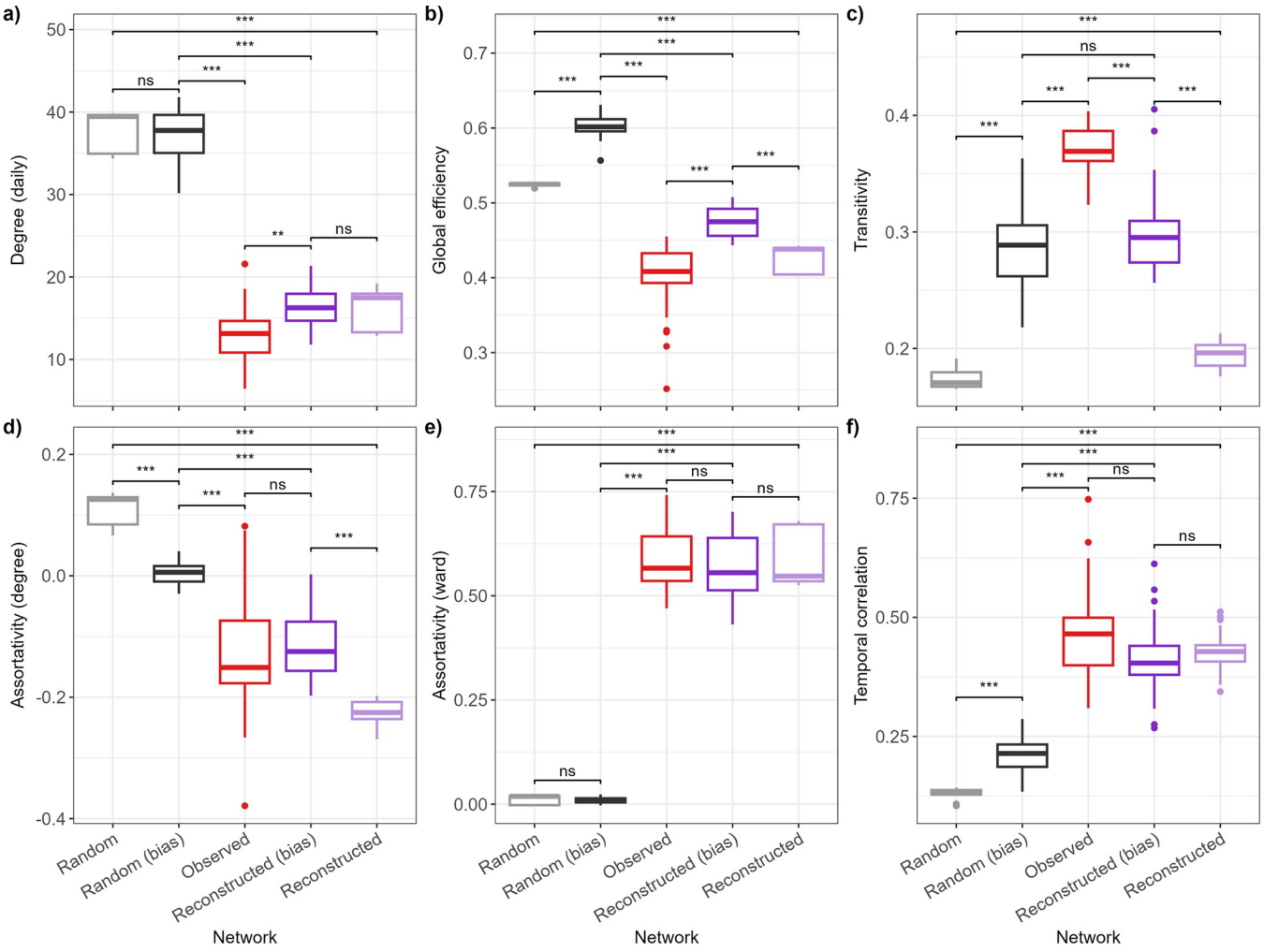


Fig 3. Comparison of network characteristics. The reconstructed networks with observation bias exclude individuals from the network at times when they were known to not wear their sensors. The random networks did not take into account the ward-level structure of the contacts or the probability of recurring 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 distributions were compared using Wilcoxon tests with Bonferroni correction. ns: not significant; *: $p < 0.05$; **: $p < 0.01$; ***: $p < 0.001$.

https://doi.org/10.1371/journal.pcbi.1012227.g003

present in the reconstructed network without bias, leading to approximately twice as many contacts in that network (Fig 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 (Fig 4A).

The distributions of contact durations in the synthetic networks were similar to the distribution in the observed network, although there were slightly less contacts with short durations (Fig 4B, all distributions are significantly different; Wilcoxon test with Bonferroni correction p values < 0.001). This is because all networks sample their contact durations from a lognormal distribution parameterised by the mean and variance estimated from the data, which puts less emphasis on very short contacts of less than one minute (S5 Fig).

In supplementary analyses, we assessed the robustness of our algorithm by quantifying the variability of network characteristics across 100 reconstructed networks without bias (S6 Fig). The variability across reconstructed networks was not statistically significant for any metric

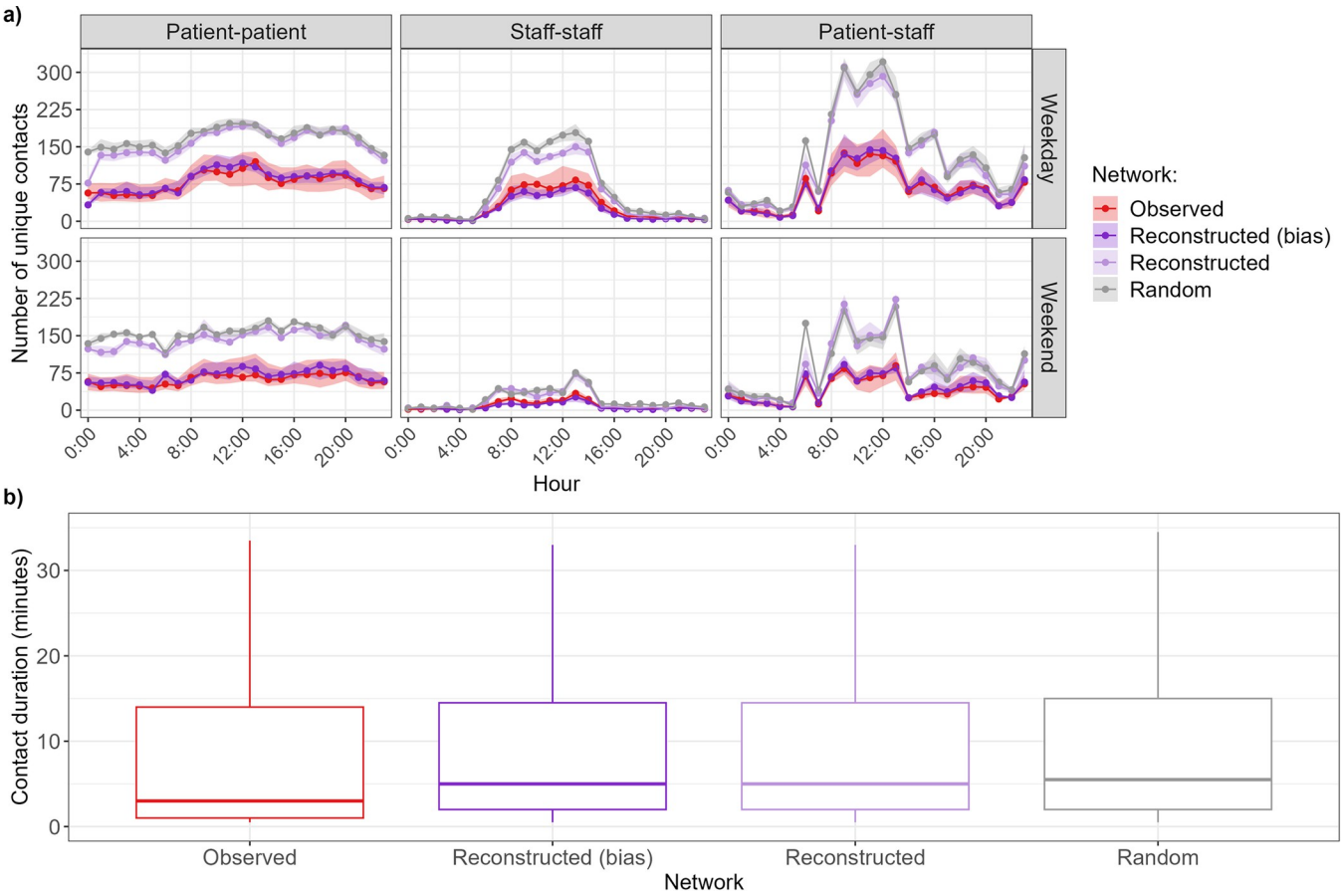
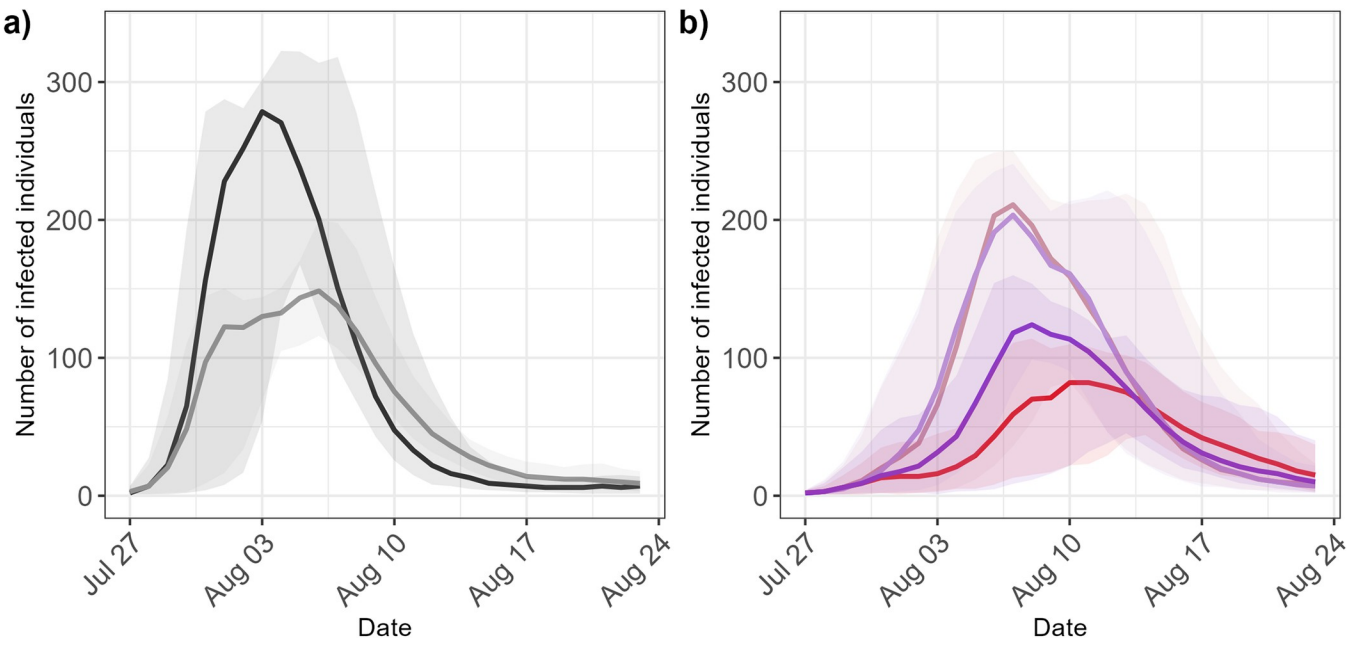


Fig 4. Comparison of network contact number and duration. a) Distribution of number of unique contacts per hour, separated by type of day (weekday or weekend). Points correspond to the median, and the shaded areas correspond to the interquartile range. **b) Distribution of contact durations.** For ease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

https://doi.org/10.1371/journal.pcbi.1012227.g004

(Kruskal-Wallis test, p value > 0.05) except for assortativity by degree ($p < 0.001$). We also aimed to validate our approach by generating "re-simulated" networks informed by summary statistics derived from the reconstructed networks with bias. These re-simulated networks are similar to the full reconstructed networks, indicating that our algorithm consistently recreates realistic networks and reconstructs unobserved contacts (S7 Fig). However, the number of patient-patient contacts in the re-simulated networks is slightly higher than in the reconstructed networks (S7 Fig).

Finally, we simulated epidemics on the networks, and found that, despite constant transmission parameters, epidemic dynamics varied depending on which network was used (Fig 5). The random networks generated the fastest and largest epidemics, with a median final attack rate of 0.91 (interquartile range: 0.89–0.93), which corresponds to the proportion of all individuals infected by the end of the one-month period. Extinctions, defined as epidemics with less than 10 individuals infected in total, never occurred on the random network, while they occurred in 17% of simulations on the observed network. Random or reconstructed networks with bias led to smaller epidemics than their non-biased counterparts, with a higher risk of extinction. The reconstructed and re-simulated networks generated similar epidemic curves. We repeated these simulations, changing the average duration of infectiousness to 2 or 10 instead of 5 days (corresponding to basic reproduction numbers of 1.3 or 6.5 instead of 3.25).



c)

Network	Median peak date	Median peak (IQR)	Median attack rate (IQR)	Extinction prop.
Random	2009-08-03	290 (266-329)	0.91 (0.89-0.93)	0.00
Random (bias)	2009-08-06	152 (134-200)	0.79 (0.76-0.82)	0.06
Observed	2009-08-10	87 (41-117)	0.5 (0.28-0.6)	0.17
Reconstructed (bias)	2009-08-08	128 (77-160)	0.63 (0.38-0.71)	0.14
Reconstructed	2009-08-07	214 (180-241)	0.83 (0.79-0.86)	0.06
Re-simulated	2009-08-07	224 (179-255)	0.84 (0.78-0.87)	0.03

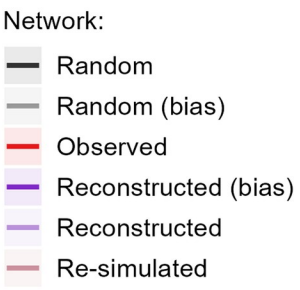


Fig 5. Comparison of resulting incidence dynamics depending on the networks. a) Epidemic dynamics for the two random networks. Lines indicate median values, and the shaded areas indicate the interquartile range. **b) Epidemic dynamics for the observed and reconstructed networks.** Lines indicate median values, and the shaded areas indicate the interquartile range. **c) Characteristics of the resulting epidemics for the different evaluated networks.** Peak date is shown for the median epidemic curve (solid lines in panels a and b). "Extinction prop." indicates the proportion of simulations excluded from the analysis, with less than 10 individuals infected in total.

https://doi.org/10.1371/journal.pcbi.1012227.g005

These changes respectively led to larger and smaller epidemics, but did not affect our qualitative conclusions (S8 Fig).

Discussion

Summary of findings

In this article, we present an approach to construct stochastic synthetic temporal contact networks in HCS from partially observed contact data. The i-Bird network illustrates the typical complex contact structures in HCS, notably with 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. Our reconstruction algorithm successfully captured the heterogeneity of the observed network by taking into account contact rates by hour, type of day (weekday or weekend) and staff category, and probabilities of recurring contacts estimated for patients and staff. The resulting reconstructed networks reproduced well the characteristics of the observed network, as well as the specific distribution of unique contacts per hour. We have previously shown that, in the i-Bird network, ward and staff category are associated with the contact frequency and duration of individuals [31]. Our findings echo this point, since the pseudo-random networks which did not consider contact heterogeneities by ward and staff category did not possess the same characteristics as the observed network. Lastly, predicted epidemic dynamics varied depending on which network they were simulated on. Although a fully observed ground-truth network was not available for comparison, the resulting epidemics were larger in the full reconstructed networks compared to the biased ones, highlighting the impact of missing contacts on disease transmission.

The value of approaches to stochastically generate realistic contact networks has been previously discussed for schools or workplaces [6,34], although the complexity of the contact structures in those settings is arguably lower than what we observed here. These approaches extended networks by repeating contact structures at fixed intervals, either keeping all nodes and links identical, or by introducing some stochasticity by randomly changing the identity of nodes (i.e. individuals) each day [6,34]. On the other hand, our algorithm dynamically and stochastically constructs new contacts at each hour based on the empirical contact rates, instead of repeating links. Our algorithm extends the principle of the stochastic block model, which relies on dividing individuals into a fixed number of groups and determining between- and within-group contact rates [35]. Here, we applied this principle across multiple overlapping categories (patient or staff categories, and ward), whilst additionally considering temporal heterogeneity (daily, hourly, and for the probability of recurring contacts). Previous algorithms also attempted to reconstruct missing contacts for non-participants [36]. While this was not accounted for here (e.g. visitors, see below for details), here we conduct this reconstruction at a higher resolution, since in reality participating individuals can also have contact data missing only for some hours or days of their total presence time. Finally, the novelty of our approach here is that we conduct a direct comparison between the output of our algorithm and the observed contact network, as opposed to other algorithms which attempted to build networks directly from contact diaries and hence did not have access to an observed network for comparison [37].

Similarities between the observed and reconstructed networks

Although 90% of individuals agreed to wear a sensor during the study, the i-Bird contact network was only partially observed, since the median time when contacts were not recorded was 33.3% (IQR: 10.5–53.6%) of a patient's presence days (40.0%, IQR: 0–75.0% for staff). This could have occurred for a number of reasons which we cannot distinguish, including depleted batteries, sensor malfunction, imperfect sensor-wearing compliance, or temporary patient releases from the facility (see Limitations below). 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 (S2 Fig), it can be assumed that contact patterns during unobserved times were similar to those on observed times. 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 (Figs 5 and S7), demonstrating the consistency of our algorithm to reconstruct contacts.

The 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 (S6 Fig). 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 [38–41]. 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 may be an important factor to explain why simulated epidemics were substantially faster and larger on random networks than reconstructed ones (Fig 5). 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 reducing the number of nodes available in the network. Simultaneously, there was still a need at those times to generate some novel contacts between individuals who never previously met, since the probability of recurring contacts was less than 1. Combined, these elements increased the overall connectivity amongst all individuals in the reconstructed network with bias. Although this could facilitate disease transmission across these reconstructed networks if they are used for such purpose [42], the high assortativity by ward may counter this effect by slowing down transmission across the entire healthcare facility.

Our algorithm did not specifically account for transitivity when recreating contacts. This is likely why the resulting transitivity was similar to that of the random network and underestimated the observed value (Fig 3). Similarly to density and global efficiency mentioned above, any transitivity in the reconstructed network was likely an indirect consequence of assortativity by ward, restricting the pool of available individuals to generate contacts and leading to interconnectivity between individuals present in the same ward. Whilst we could extend our algorithm to consider transitivity when choosing the individuals to put in contact, we decided not to do this here to maximise the generalisability of our approach by not requiring such highly detailed contact data. In any case, this may not substantially affect disease transmission simulated across these networks, since previous work has shown that transitivity is a poor predictor of the total number of individuals who would be infected across the network [42].

Although our algorithm can capture individual presence and absence times, information about patient temporary releases from the LTCF (e.g., for weekends with their families, or for

shopping outside) was not available in the i-Bird data, hence such events were not accounted for here, although they may occur frequently in a LTCF. Consequently, the number of presence days/hours may have been overestimated, leading to an overestimation of contact days among patients. Although this is negligible when comparing the observed and reconstructed network with bias, this is likely why the re-simulated networks slightly overestimated the number of patient-patient contacts compared to the full reconstructed network (S7 Fig). We expect that this overestimation would be absent in HCS with more complete information on individual presence, or in acute care facilities with shorter patient lengths of stay and where temporary releases are less common. Similarly, our algorithm does not consider the contacts of visitors in the hospital, and we did not have data in the i-Bird study on visitors which we would have required to validate the synthetic networks. Consequently, our description of the contact structure in the LTCF is not exhaustive, although this does not affect the ability of our algorithm to reproduce patient-staff contacts.

When reconstructing missing contacts, we assumed that if a staff member (patient) was present in the facility at a given time but did not have any contact with anyone else recorded at that hour (day), this represented unobserved data. In reality, there may be rare instances where individuals truly did not have any contact with anyone else over a time period. In such instances, our algorithm would over-estimate contacts by forcibly reconstructing contacts for those individuals at those times. However, we expect this would only occur at times with limited contact rates (e.g. during the night), therefore the empirical contact rates would be small and only a couple of contacts may be erroneously reconstructed by the algorithm.

More broadly, our approach to reconstruct missing contacts relies on the assumption that, for a given set of individual characteristics, unobserved contacts follow the same distribution as observed ones. For example, we assume that the unobserved contacts of a nurse *n1* in ward *w* at hour *h* are similar to the observed contacts of another nurse *n2* also in ward *w* at hour *h*. We found this assumption to be reasonable in the observed i-Bird network, since the average contact rates of individuals did not correlate with the amount of time when they had unobserved contacts (S2 Fig). However, this assumption would have to be carefully checked when attempting to use our algorithm with other data sources or other types of settings.

Future work

In this study, we show that our algorithm can accurately reproduce the contact structure using as input contact data from a given long-term care facility. Although our algorithm has not been designed to reproduce mobility networks which possess different characteristics, it should be applicable to any contact network representing close proximity interaction between individuals in a defined setting. A first important next step would be to repeat our analysis using data collected in a different HCS such as acute care, over a different time period. This is because contact structures are known to vary between different HCS such as long-term or acute, with more/less contacts between different individual categories, varying recurring contact probabilities etc. Similarly, even though we tested our algorithm using substantial data covering four weeks, this contact structure may not be representative of other time periods. Notably, the i-Bird data we used was collected in the middle of the summer, which is a holiday period in France and may have affected contact patterns. Although we do not expect that our algorithm will perform differently since it has been designed to be generalisable, 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 while maintaining constrains on each population size. This would allow us to further account for possible variability in the structure of the population in the facility, add flexibility in building synthetic networks for settings where this data may not be fully available, and hence add further stochasticity in our algorithm.

Since contact data may only be available for short periods of time (e.g. a few days [22]), 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 (S9 Fig). As expected, this led to variability amongst the reconstructed networks depending on the length of the period used, since this reduced the number of data points used to estimate the average contact rates used by the algorithm. In any case, the main risk of using only a short period of time is to miss out on 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. A further extension of our algorithm could include the possibility of creating such unobserved links, but this would still require either assumptions or information on the nature of those links. 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 in the studied LTCF over the studied period as 0.71 for staff and 0.78 for patients, but we note some individual variation in this value (Fig 2, interquartile range for staff: 0.63–0.84, for patients: 0.77–0.85). In addition, our estimation here is made using the entire observed contact networks over the study period, 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 (S10 Fig). 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 key to better understand contact formation and heterogeneity, and could be directly taken into consideration in our algorithm since it has been designed to use this probability. In healthcare settings, this probability could likely be estimated without requiring complete contact data, using information on staff schedules and patient ward or room allocation instead.

Finally, other methodological approaches could be considered to reconstruct realistic contact networks. For example, deep learning algorithms such as graph convolutional networks (GCN) have become increasingly popular for this purpose, particularly in the context of infectious disease transmission [43–47]. 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 [48,49]. On the other hand, temporal graph networks can capture this temporal dependency [50,51], 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 require large amounts of training data. Democratising their use would therefore first require new studies to collect close-proximity interaction data in different settings and time periods, presenting further logistical challenges.

Implications

Our algorithm relies on computing summary statistics from an observed network, then using these statistics to stochastically reconstruct contact networks. Such statistics can be derived directly from other observed networks, as we have done here to validate our approach. In that case, instead of only relying on a single observed network, our approach provides multiple realistic reconstructed networks enabling to consider the impact of stochasticity of the contact structure and on subsequent epidemic risk in a given setting. Our approach, by providing data augmentation, also enables to infer information on potentially unobserved contacts. Our algorithm could also be used to generate extended realistic temporal dynamics over longer time periods than the period of data collection. However, as we have shown in our supplementary analysis using only part of the data to estimate contact parameters, this must be further studied to clarify the minimal amount of data required to be confident in the representativity of the synthetic networks.

Alternatively, summary contact statistics could be more simply collected from cross-sectional surveys or even derived exclusively from individual schedules, which would not require a detailed and costly follow-up using sensors. In this scenario, the only other data required would be individual presence times, which should either be routinely available (e.g. in healthcare settings or schools) or relatively easy to collect (e.g. in workplaces). Although as mentioned in the Limitations, the amount of data our algorithm requires to generate realistic networks is still unclear, our approach could ultimately be used to generate contact networks from contact matrices. This would substantially facilitate research on the impact of contact heterogeneity in various populations and settings, as others have previously discussed [37].

In conclusion, our algorithm can generate temporal contact networks in a healthcare setting by taking into consideration empirically measured contact rates based on close-proximity sensors, as opposed to most available packages which only construct static networks and rely on hyperparameters [52,53]. These temporal networks can then be analysed with mathematical models to evaluate the potential impact of interventions against disease transmission [11–14]. In particular, this will improve the wider applicability of individual-based model which make it possible to account for detailed contact heterogeneity in testing the effect of interventions targeting highly specific individuals.

Supporting information

S1 Table. Summary of network characteristics for the observed total network, patient-patient subgraph, staff-staff subgraph, and patient-staff subgraph, separated by weekday or weekend. Values were estimated per day, with mean and standard deviation (sd) presented here. Transitivity is not shown for the patient-staff subgraph as triangles of contacts cannot occur in this network.
(DOCX)

S1 Fig. Proportion of presence days during which contact data were recorded, by patient ward and staff category. Each point is one individual's proportion, calculated over the entire study period. The shading of the points indicate the number of presence days for each individual (darker = more presence days).
(TIF)

S2 Fig. Proportion of presence days during which contact data were recorded is not correlated with a) average number of contacts per day nor b) number of presence days. Each point is one individual.
(TIF)

S3 Fig. Temporal correlation by subgraph and day of the week. The correlation is calculated for each day by comparing it to the previous day; for example, the value on "Monday" indicates the correlation between the network on Monday and Sunday. Points indicate the mean correlation, and lines indicate the 95% confidence interval (1.96 times the standard deviation). Weekdays are shown in blue, and weekends in red.
(TIF)

S4 Fig. Comparison of density across networks. 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.
(TIF)

S5 Fig. Contact rates in the i-Bird data and sampled from a lognormal distribution. 100,000 samples are taken from a lognormal distribution is informed by the mean and variance estimated from the data. For ease of visualisation, the x-axis is truncated at 60 minutes.
(TIF)

S6 Fig. Variability of network characteristics across iterations. Here, 100 reconstructed networks without bias were generated independently, all informed by the same contact rates estimated from the i-Bird data. For each iteration, the distribution of the metric calculated for each day is shown. Red points are weekdays, and blue points are weekends. The green lines indicate the median value for the correspond metric across all networks and days. Only the distributions of assortativity by degree are significantly different between networks (Kruskal-Wallis test, p value < 0.001).
(TIF)

S7 Fig. Full reconstructed networks informed by the observed network, compared to full re-simulated networks informed by the reconstructed network with observation bias.
(TIF)

S8 Fig. Comparison of resulting incidence dynamics depending on the networks, with an average duration of infectiousness of 2 days (a-b) or 10 days (c-d). a-c) Epidemic dynamics for the two random networks. Lines indicate median values, and the shaded areas indicate the interquartile range. **b-d) Epidemic dynamics for the observed and reconstructed networks.** Lines indicate median values, and the shaded areas indicate the interquartile range.
(TIF)

S9 Fig. Characteristics for the full reconstructed network, estimated using two weeks or one week instead of the complete observed four weeks. 1st quarter: 26/07–02/08, 2nd quarter:

03/08–09/08, 3rd quarter: 10/08–16/08, 4th quarter: 17/08–23/08. "1st half" includes the 1st and 2nd quarters, and "2nd half" includes the 3rd and 4th quarters.
(TIF)

S10 Fig. Characteristics for the full reconstructed network when manually setting recurring contact probability to 0.1, 0.5 or 0.9. The calculated probabilities are 0.78 for patients and 0.71 for staff. In the other scenarios, the probabilities for patients and staff are set to the same value.
(TIF)

Acknowledgments

The authors would like to thank Eric Fleury, Pierre-Yves Boëlle, Vittoria Colizza and Pascal Crépey for helpful discussions on the analysis of the contact network.

Author Contributions

Conceptualization: Audrey Duval, Quentin J. Leclerc, Didier Guillemot, Laura Temime, Lulla Opatowski.

Formal analysis: Audrey Duval, Quentin J. Leclerc.

Funding acquisition: Didier Guillemot.

Investigation: Audrey Duval, Quentin J. Leclerc, Laura Temime, Lulla Opatowski.

Methodology: Audrey Duval, Quentin J. Leclerc, Laura Temime, Lulla Opatowski.

Software: Audrey Duval, Quentin J. Leclerc.

Supervision: Didier Guillemot, Laura Temime, Lulla Opatowski.

Validation: Didier Guillemot, Laura Temime, Lulla Opatowski.

Visualization: Quentin J. Leclerc.

Writing – original draft: Audrey Duval, Quentin J. Leclerc.

Writing – review & editing: Audrey Duval, Quentin J. Leclerc, Didier Guillemot, Laura Temime, Lulla Opatowski.

References

1. Keeling MJ, Rohani P. Modeling Infectious Diseases in Humans and Animals. Modeling Infectious Diseases in Humans and Animals. Princeton University Press; 2008. https://doi.org/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. https://doi.org/10.1371/journal.pone.0250050 PMID: 34283842
5. Machens A, Gesualdo F, Rizzo C, Tozzi AE, Barrat A, Cattuto C. An infectious disease model on empirical networks of human contact: bridging the gap between dynamic network data and contact matrices. BMC Infectious Diseases. 2013; 13: 185. https://doi.org/10.1186/1471-2334-13-185 PMID: 23618005
6. Stehlé J, Voirin N, Barrat A, Cattuto C, Colizza V, Isella L, et al. Simulation of an SEIR infectious disease model on the dynamic contact network of conference attendees. BMC Medicine. 2011; 9: 87. https://doi.org/10.1186/1741-7015-9-87 PMID: 21771290

7. Kiss IZ, Miller JC, Simon PL, others. Mathematics of epidemics on networks. Cham: Springer. 2017; 598: 31.
8. 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. https://doi.org/10.1155/2011/284909 PMID: 21437001
9. 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. https://doi.org/10.1371/journal.pmed.0050074 PMID: 18366252
10. Keeling MJ, Eames KTD. Networks and epidemic models. Journal of The Royal Society Interface. 2005; 2: 295–307. https://doi.org/10.1098/rsif.2005.0051 PMID: 16849187
11. 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. https://doi.org/10.1080/17513758.2010.503376 PMID: 22877143
12. Masuda N, Holme P. Predicting and controlling infectious disease epidemics using temporal networks. F1000Prime Rep. 2013; 5: 6. https://doi.org/10.12703/P5-6 PMID: 23513178
13. Gross T, D'Lima CJD, Blasius B. Epidemic Dynamics on an Adaptive Network. Phys Rev Lett. 2006; 96: 208701. https://doi.org/10.1103/PhysRevLett.96.208701 PMID: 16803215
14. 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. https://doi.org/10.1371/journal.pcbi.1004152 PMID: 25763816
15. Holme P, Saramäki J. Temporal networks. Physics Reports. 2012; 519: 97–125. https://doi.org/10.1016/j.physrep.2012.03.001
16. 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. https://doi.org/10.1093/infdis/jis542 PMID: 23045621
17. 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. https://doi.org/10.1371/journal.pcbi.1004170 PMID: 25789632
18. Salathé M, Kazandjieva M, Lee JW, Levis P, Feldman MW, Jones JH. A high-resolution human contact network for infectious disease transmission. Proceedings of the National Academy of Sciences. 2010; 107: 22020–22025. https://doi.org/10.1073/pnas.1009094108 PMID: 21149721
19. 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. https://doi.org/10.1016/j.compbiomed.2021.104787 PMID: 34482197
20. Eames K, Bansal S, Frost S, Riley S. Six challenges in measuring contact networks for use in modelling. Epidemics. 2015; 10: 72–77. https://doi.org/10.1016/j.epidem.2014.08.006 PMID: 25843388
21. 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. https://doi.org/10.1017/S0950268812000842 PMID: 22687447
22. Vanhems P, Barrat A, Cattuto C, Pinton J-F, Khanafer N, Régis C, et al. Estimating Potential Infection Transmission Routes in Hospital Wards Using Wearable Proximity Sensors. Viboud C, editor. PLoS ONE. 2013; 8: e73970. https://doi.org/10.1371/journal.pone.0073970 PMID: 24040129
23. Stehlé J, Voirin N, Barrat A, Cattuto C, Isella L, Pinton J-F, et al. High-Resolution Measurements of Face-to-Face Contact Patterns in a Primary School. PLOS ONE. 2011; 6: e23176. https://doi.org/10.1371/journal.pone.0023176 PMID: 21858018
24. 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. https://doi.org/10.1186/s12879-016-1676-y PMID: 27449511
25. Watts DJ, Strogatz SH. Collective dynamics of 'small-world' networks. Nature. 1998; 393: 440–442. https://doi.org/10.1038/30918 PMID: 9623998
26. Almutiry W, Deardon R. Contact network uncertainty in individual level models of infectious disease transmission. Stat Commun Infect Dis. 2021; 13: 20190012. https://doi.org/10.1515/scid-2019-0012 PMID: 35880993
27. Shirley MDF, Rushton SP. The impacts of network topology on disease spread. Ecological Complexity. 2005; 2: 287–299. https://doi.org/10.1016/j.ecocom.2005.04.005
28. 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. https://doi.org/10.1371/journal.pmed.1003907 PMID: 35231023

29. 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. https://doi.org/10.7554/eLife.70294 PMID: 34821551

30. 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. https://doi.org/10.1097/EDE.0000000000001047 PMID: 31274572

31. 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. https://doi.org/10.1038/s41598-018-20008-w PMID: 29374222

32. 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/

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

34. Colosi E, Bassignana G, Contreras DA, Poirier C, Boëlle P-Y, Cauchemez S, et al. Screening and vaccination against COVID-19 to minimise school closure: a modelling study. The Lancet Infectious Diseases. 2022; 22: 977–989. https://doi.org/10.1016/S1473-3099(22)00138-4 PMID: 35378075

35. Lee C, Wilkinson DJ. A review of stochastic block models and extensions for graph clustering. Appl Netw Sci. 2019; 4: 1–50. https://doi.org/10.1007/s41109-019-0232-2

36. Génois M, Vestergaard CL, Cattuto C, Barrat A. Compensating for population sampling in simulations of epidemic spread on temporal contact networks. Nat Commun. 2015; 6: 8860. https://doi.org/10.1038/ncomms9860 PMID: 26563418

37. Mastrandrea R, Barrat A. How to Estimate Epidemic Risk from Incomplete Contact Diaries Data? PLOS Computational Biology. 2016; 12: e1005002. https://doi.org/10.1371/journal.pcbi.1005002 PMID: 27341027

38. Tang J, Scellato S, Musolesi M, Mascolo C, Latora V. Small-world behavior in time-varying graphs. Phys Rev E. 2010; 81: 055101. https://doi.org/10.1103/PhysRevE.81.055101 PMID: 20866285

39. Kretzschmar M, Morris M. Measures of concurrency in networks and the spread of infectious disease. Mathematical Biosciences. 1996; 133: 165–195. https://doi.org/10.1016/0025-5564(95)00093-3 PMID: 8718707

40. Read JM, Eames KTD, Edmunds WJ. Dynamic social networks and the implications for the spread of infectious disease. J R Soc Interface. 2008; 5: 1001–1007. https://doi.org/10.1098/rsif.2008.0013 PMID: 18319209

41. Smieszek T, Fiebig L, Scholz RW. Models of epidemics: when contact repetition and clustering should be included. Theor Biol Med Model. 2009; 6: 11. https://doi.org/10.1186/1742-4682-6-11 PMID: 19563624

42. 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. https://doi.org/10.1038/s41598-022-11786-5 PMID: 35545647

43. 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. https://doi.org/10.1038/s41598-022-07757-5 PMID: 35273252

44. 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. https://doi.org/10.1093/jamia/ocaa322 PMID: 33486527

45. 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. https://doi.org/10.1609/aaai.v35i6.16616

46. 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. https://doi.org/10.48550/arXiv.2007.03113

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

48. 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. https://doi.org/10.1016/j.jbi.2021.103720 PMID: 33640536

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

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

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

52. Prettejohn B, Berryman M, McDonnell M. Methods for Generating Complex Networks with Selected Structural Properties for Simulations: A Review and Tutorial for Neuroscientists. Frontiers in Computational Neuroscience. 2011; 5. Available: https://www.frontiersin.org/articles/10.3389/fncom.2011.00011 PMID: 21441986

53. Hagberg AA, Schult DA, Swart PJ. Exploring Network Structure, Dynamics, and Function using NetworkX. In: Varoquaux G, Vaught T, Millman J, editors. Proceedings of the 7th Python in Science Conference. Pasadena, CA USA; 2008. pp. 11–15.