

1 An algorithm to build synthetic temporal contact 2 networks based on close-proximity interactions data

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

4 **Audrey Duval^{1,2,3+✕}, Quentin Leclerc^{1,2,3+*}, Didier Guillemot^{1,2,4}, Laura Temime^{3,5#}, Lulla
5 Opatowski^{1,2#}**

6 * corresponding author (quentin.leclerc@pasteur.fr)

7 + these authors contributed equally

8 # these authors contributed equally

9

10 ¹ Institut Pasteur, Université Paris Cité, Epidemiology and Modelling of Bacterial Escape to
11 Antimicrobials (EMEA), 75015 Paris, France

12 ² INSERM, Université Paris-Saclay, Université de Versailles St-Quentin-en-Yvelines, Team
13 Echappement aux Anti-infectieux et Pharmacoépidémiologie U1018, CESP, 78000 Versailles,
14 France

15 ³ Laboratoire Modélisation, Epidémiologie et Surveillance des Risques Sanitaires (MESuRS),
16 Conservatoire National des Arts et Métiers, 73003 Paris, France

17 ⁴ AP-HP, Paris Saclay, Department of Public Health, Medical Information, Clinical research, F-
18 92380, Garches

19 ⁵ Institut Pasteur, Conservatoire National des Arts et Métiers, Unité PACRI, 75015 Paris, France

20

21 ✕ Current address : Imagine Institute, Data Science Platform, INSERM UMR 1163, Université
22 de Paris, Paris, France

23

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

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.

26 **Acknowledgments**

27 AD, LT and LO received funding from the French National Research Agency (SPHINX-17-CE36-
28 0008-01). DG received funding from the National Clinical Research Program and the
29 Investissement d'Avenir program, Laboratoire d'Excellence "Integrative Biology of Emerging
30 Infectious Diseases" (ANR-10-LABX-62-IBEID). The authors would like to thank Eric Fleury,
31 Pierre-Yves Boëlle, Vittoria Colizza and Pascal Crépey for helpful discussions on the analysis of
32 the contact network.

33

34 **Abstract**

35 Small populations (e.g., hospitals, schools or workplaces) are characterised by high contact
36 heterogeneity and stochasticity affecting pathogen transmission dynamics. Empirical
37 individual contact data provide unprecedented information to characterize such
38 heterogeneity and are increasingly available, but are usually collected over a limited period,
39 and can suffer from observation bias. We propose an algorithm to stochastically reconstruct
40 realistic temporal networks from individual contact data in healthcare settings (HCS) and test
41 this approach using real data previously collected in a long-term care facility (LTCF).

42 Our algorithm generates full networks from recorded close-proximity interactions, using
43 hourly inter-individual contact rates and information on individuals' wards, the categories of
44 staff involved in contacts, and the frequency of recurring contacts. It also provides data
45 augmentation by reconstructing contacts for days when some individuals are present in the
46 HCS without having contacts recorded in the empirical data. Recording bias is formalized
47 through an observation model, to allow direct comparison between the augmented and
48 observed networks. We validate our algorithm using data collected during the i-Bird study,
49 and compare the empirical and reconstructed networks.

50 The algorithm was substantially more accurate to reproduce network characteristics than
51 random graphs. The reconstructed networks reproduced well the assortativity by ward (first–
52 third quartiles observed: 0.54–0.64; synthetic: 0.52–0.64) and the hourly staff and patient
53 contact patterns. Importantly, the observed temporal correlation was also well reproduced

54 (0.39–0.50 vs 0.37–0.44), indicating that our algorithm could recreate a realistic temporal
55 structure. The algorithm consistently recreated unobserved contacts to generate full
56 reconstructed networks for the LTCF.

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

62

63 **Author summary**

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

79 Introduction

80 Limiting the burden of infectious diseases requires a good understanding of how they spread.
81 For pathogens transmitted mostly via close-proximity interactions, the rate at which
82 individuals come into contact with each other is strongly correlated with the expected spread
83 of the disease across the population [1]. In large populations such as cities or countries,
84 contact structures are usually approximated by grouping individuals into relatively broad
85 categories (neighbourhood, age...), and assuming that contact rates are heterogeneous
86 between categories, but homogeneous within [2,3]. In small populations such as healthcare
87 institutions, schools, or workplaces however, disease transmission is affected by high contact
88 heterogeneity and stochasticity [4]. Capturing these characteristics requires a detailed,
89 individual-level description of contacts instead of only relying on summary contact rates by
90 groups [5,6].

91 Contact networks are increasingly used to fully capture the interactions between individuals
92 in small populations [7,8]. These networks explicitly represent the links between all individuals
93 in such populations, as opposed to contact matrices which capture average contact rates
94 between groups of individuals [9,10]. Temporal contact networks further capture the time-
95 changing nature of contacts, therefore representing individual interactions more accurately
96 than static networks [11–15]. Contact networks can be coupled with individual-based
97 mathematical models to help design effective interventions against the spread of infectious
98 diseases, since they enable the identification of highly connected individuals who can be
99 targeted to lead to the greatest impact on transmission [10]. Recently, empirical data
100 collected to build inter-individual temporal networks has become increasingly available to
101 inform contact networks. For example, studies have used sensors to record close-proximity
102 interactions between individuals [16–18], and contact tracing programs have relied on the
103 integrated Bluetooth technology in mobile phones [19].

104 However, the detailed empirical data required to build temporal contact networks remain
105 subject to several limitations [20,21]. These data are typically collected over a few days only
106 [22,23], and may be subject to observation bias; sensors might not be properly placed to
107 register contacts [24], or individuals may disable Bluetooth on their mobile phones at different
108 times [19]. Due to the resulting missed contacts, the networks derived from these data may

109 only be partially observed. Transmission rates estimated using these partially observed
110 networks would be overestimated compared to reality due to the lower number of contacts,
111 which could lead to an incorrect evaluation of the impact of interventions [25–27]. By
112 comparison, although they do not provide individual-level information, contact matrices and
113 summary statistics such as contact rates between individual groups are more readily available,
114 as they can be inferred using simple cross-sectional survey data [28–30].

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

122 **Methods**

123 **Building synthetic contacts in a HCS**

124 *Algorithm outline*

125 We built an algorithm to stochastically reconstruct a realistic full temporal network of inter-
126 individual close-proximity interactions (CPIs, at less than 1.5m) in a HCS using parameters
127 estimated from empirical individual contact data. This algorithm generates a new synthetic
128 network which notably reconstructs contacts at times when individuals were known to be
129 present in the HCS but had no contact data recorded, which we consider to be a recording
130 bias. The synthetic network hence includes both the observed and unobserved parts of the
131 empiric network. This approach first involves the calculation of contact rates and durations
132 between individuals, stratified by the individuals' ward, category (patient, or staff profession),
133 type of day (weekday or weekend) and hour. The algorithm then reconstructs a new network,
134 taking as input these summary statistics as well as data on presence days for each individual
135 in the facility. Each CPI is generated stochastically, with individuals chosen in order to promote
136 recurring contacts, based on a probability estimated from the data.

137

138 *Estimation of contact rates from the data*

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

$$141 \quad T_{h,c_1w_1 \rightarrow c_2w_2} = \frac{\sum_{i \in C_1 W_1} \sum_{j \in C_2 W_2} \sum_{k=1}^{N_{h,i}} V_{i,j,k}}{\sum_{l=1}^{N_h} N_{C_1 W_1, l}} \quad (1)$$

142 where $T_{h,c_1w_1 \rightarrow c_2w_2}$ is the average per-person contact rate at the hour h between individuals
143 from category C_1 belonging to ward W_1 and individuals from category C_2 belonging to ward
144 W_2 . For given hour h and individual i , $N_{h,i}$ is the number of instances of the hour h where at
145 least one contact was recorded for individual i . For example, if i had a contact recorded on
146 Tuesday 11th August at 10h, and on Tuesday 18th August at 10h, $N_{10,i}$ would be equal to 2. For
147 two individuals i from $C_1 W_1$ and j from $C_2 W_2$, $V_{i,j,k}$ indicates whether contacts have been
148 recorded between them on instance k of the hour h : it equals 1 if i and j had at least one
149 contact recorded at that time, and 0 otherwise. Finally, N_h is the total number of instances of

150 the hour h in the full dataset and, for a given instance l of the hour h , $N_{C_1W_1,l}$ is the number of
151 individuals from C_1W_1 that had any contact recorded during that hour.

152

153 This estimation is conducted separately for contacts during weekdays and contacts during
154 weekends.

155

156 ***Estimation of recurring contacts***

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

$$159 \quad p_{i,d} = \frac{|U_{i,d} \cap U_{i,[d_0,d]}|}{|U_{i,d}|} \quad (2)$$

160 Where $U_{i,d}$ is the set of unique individuals with whom i had a contact on day d , $U_{i,[d_0,d]}$ is the
161 set of unique individuals with whom i had at least one contact on any day between the first
162 day d_0 and the current day d (d non-included), and the notation $|x|$ indicates the cardinal of
163 the set x . For example, if i had a contact with four unique individuals on day d , and previously
164 had a contact with two of those on any day between d_0 and d , the probability of recurring
165 contact for day $p_{i,d}$ would be $2/4 = 0.5$.

166

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

$$169 \quad p_i = \frac{\sum_{d=d_0}^{d_{max}} p_{i,d}}{1+(d_{max}-d_0)} \quad (3)$$

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

$$172 \quad p_c = \frac{\sum_{i \in C} p_i}{|C|} \quad (4)$$

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

174

175 ***Generation of synthetic CPIs: number and identity of individuals in contacts***

176 For each hour of our period of interest, we estimate the number of contacts between
177 individuals present in the HCS during that hour, determined using admission data and staff
178 schedule. We generate the number of individuals n from category C_2S_2 in contact with an
179 individual i from category C_1S_1 during an hour h by sampling from a Poisson distribution with

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

185
186 These individuals are selected by favouring contacts between individuals who have already
187 met at any other time previous to h . Let p_c be the probability of a recurring contact for
188 category c (patient or staff) of the individual i . To determine the identity of the n individuals
189 in contact with i , we draw a random number $r \sim Uniform(0,1)$

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

192
$$p_{i \rightarrow j} = \frac{N_{i \rightarrow j}}{\sum_{k \in S} N_{i \rightarrow k}} \quad (5)$$

193 Where $N_{i \rightarrow j}$ is the number of previous contacts between i and j before hour h , and
194 $\sum_{k \in S} N_{i \rightarrow k}$ is the number of previous contacts between i and each individual k belonging
195 to S .

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

198

199 **Generation of contact durations**

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

203

204 **Validation dataset: the i-Bird network**

205 **Dataset description**

206 We validate our algorithm by applying it to data collected during the Individual-Based
207 Investigation of Resistance Dissemination (i-Bird) study [17,31]. This study took place in a
208 rehabilitation and long-term care facility (LTCF) from the beginning of July to the end of
209 October 2009. Over this period, each participant (patient or hospital staff) was wearing an
210 RFID sensor that recorded CPIs every 30 seconds. Here, we only used contacts recorded

211 between 27 July to 23 August 2009 (included). This period corresponds to the weeks between
212 two sensor battery replacements and hence avoids interference due to loss of contact. A
213 temporal network of proximities was therefore available over 28 days with information on
214 individual ID and ward of affectation.

215 The LTCF was structured into five wards: three neurological wards, one nutritional care ward
216 and one geriatric ward. Patients were systematically linked to a ward, whilst some staff were
217 mobile and not linked to a specific ward. For the purpose of this work, we considered here
218 that mobile staff belonged to an “artificial” 6th ward, to compute contact rates according to
219 the algorithm detailed above. Staff were divided into 13 professions: administrative,
220 animation/hairdresser, logistic, hospital service agent, porter, occupational therapist,
221 physiotherapist, other rehabilitation staff, nurse, head nurse, care assistant, medical
222 student/resident, and physician. A total of 200 patients and 213 hospital staff were included
223 and had contacts recorded during the 28 days of study.

224 We used hospital staff schedules to determine the hourly presence of each staff and
225 compared these schedules to the dates and times when staff had any contact recorded. We
226 assumed that, in reality, staff would have at least one contact with any other individual during
227 any given hour of their presence time, hence if no contact was recorded for a given hour of
228 presence we considered this was missing data rather than true absence of contact. Through
229 this, we estimated that the median percentage of a staff’s total presence time when no
230 contact data was recorded was 40.0% (interquartile range (IQR): 0-75.0%). We repeated this
231 analysis for patients at the daily instead of hourly level, as we only had access to admission
232 and discharge dates for patients. We estimated that the median time when no contact data
233 was recorded was 33.3% (interquartile range (IQR): 10.5-53.6%) of a patient’s presence days.

234 Although the overall compliance was high (90% of individuals agreed to wear a sensor), there
235 was therefore substantial heterogeneity in the individual coverage of the raw i-Bird network
236 (Supplementary Figure 1). Interestingly, there was no correlation between the proportion of
237 presence time during which contact data were recorded for a given individual and their
238 average number of contacts on presence days where data were available, nor their total
239 presence time (Supplementary Figure 2).

240

241

242 ***Observation bias process***

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

251

252 **Simulations and analysis**

253 From the analysis of the i-Bird empiric network and data, we used our algorithm to generate
254 100 full synthetic reconstructed networks, and 100 reconstructed networks with observation
255 bias. For comparison, we also generated 100 pseudo-random contact networks with
256 observation bias, and 100 without. The latter networks simulate contacts without taking into
257 account the ward, staff category, and probability of recurring contact in the calculation of
258 contact rates and durations. The patient-patient, staff-staff, and patient-staff contact rates
259 are calculated as detailed in the section “Estimation of contact rates from the data”, treating
260 all staff as if they were part of the same profession, and all individuals as if they were part of
261 the same ward. At each contact, the individual encountered is therefore chosen randomly
262 from all those present in the LTCF at that time, regardless of whether or not the individual was
263 previously encountered.

264

265 We implemented the algorithm in C++ with the repast HPC 2.3.0 library. All simulations were
266 performed on the Maestro cluster hosted by the Institut Pasteur. The networks were analysed
267 in R [32], using the igraph package [33]. The relevant contact networks and analysis code are
268 available in the following GitHub repository: https://github.com/qlleclerc/network_algorithm.

269

270

271

272

273 **Validation of the full reconstructed networks**

274 For validation, we also applied the algorithm to each of the 100 reconstructed networks with
275 bias, considering them as empiric networks. This allowed us to generate 100 new full
276 reconstructed networks from fully known networks, and confirm these “re-simulated
277 networks” were similar to the full reconstructed networks generated from the observed data.

278 Results

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

280 In this section, we illustrate the typical complexity of contact structures in HCS using the i-Bird
281 network. While the algorithm makes use of data at the hourly level, in this section the contact
282 data are aggregated at the daily level, so that if two individuals have two separate contacts
283 with each other at different times of the day, this is only counted once. The contact network
284 is considered undirected, since contacts are assumed to be reciprocal. Daily-averaged contact
285 matrices built from these data are described in a previous work [31].

286
287 We first summarise the observed temporal network recorded in the LTCF during the i-Bird
288 study, comparing the total daily network and subgraphs with only patient-patient, staff-staff,
289 or patient-staff contacts (Figure 1a-d). Table 1 provides the degree, global efficiency, density,
290 transitivity, assortativity and temporal correlation of these four networks. The mean degree
291 of the total network per day is 12.99 (standard deviation: 3.53), which corresponds to the
292 average number of unique contacts per individual per day. In the subgraphs, the degree is
293 highest in the patient-staff subgraph (8.09; sd: 1.89), although we still note a relatively
294 important number of patient-patient contacts, with a degree of 5.25 (sd: 1.87) in the
295 corresponding subgraph. The distribution of individual degrees for all individuals and all days
296 across the total network is heterogeneous, with a squared coefficient of variation equal to
297 0.44 (Figure 1e). The global efficiency of the total network is 0.40 (sd: 0.05), meaning that on
298 average the shortest path between any two individuals has a distance of 2.5 (whereby the
299 shortest path between two individuals in direct contact would be of distance 1). As expected,
300 the efficiencies are lower in the subgraphs, since we remove individuals and hence increase
301 the distance between those remaining (patient-patient: 0.25 (sd: 0.08, distance: 4); staff-staff:
302 0.32 (sd: 0.10, distance: 3.1); patient-staff: 0.31 (sd: 0.05, distance: 3.2)). Densities in the total
303 network and subgraphs are relatively low (< 0.1), indicating that less than 10% of all possible
304 connections between individuals in the network are actual observed connections.

305
306 Transitivity in the total network is high (0.37; sd: 0.02), meaning that for any two individuals a
307 and b both in contact with the same third individual c , the probability that a and b are also in

308 contact is 0.37. Transitivity is also high in the patient-patient and staff-staff subgraphs, but
 309 this metric is not relevant for the patient-staff subgraph – it is impossible for a triangle of
 310 contacts to occur in this subgraph as it excludes staff-staff and patient-patient contacts by
 311 design. Assortativity by degree is negative in the total network (-0.13; sd: 0.10), indicating that
 312 highly connected individuals are more likely to be in contact with less connected individuals.
 313 It is also strongly negative in the patient-staff subgraph (-0.42; sd: 0.14), reflecting the
 314 expected disassortivity of healthcare contacts, where each staff member is in contact with
 315 multiple patients, whilst each patient is contact with relatively few staff members. In the
 316 patient-patient and staff-staff subgraphs, assortativity by degree is positive, as frequently
 317 seen in social networks.

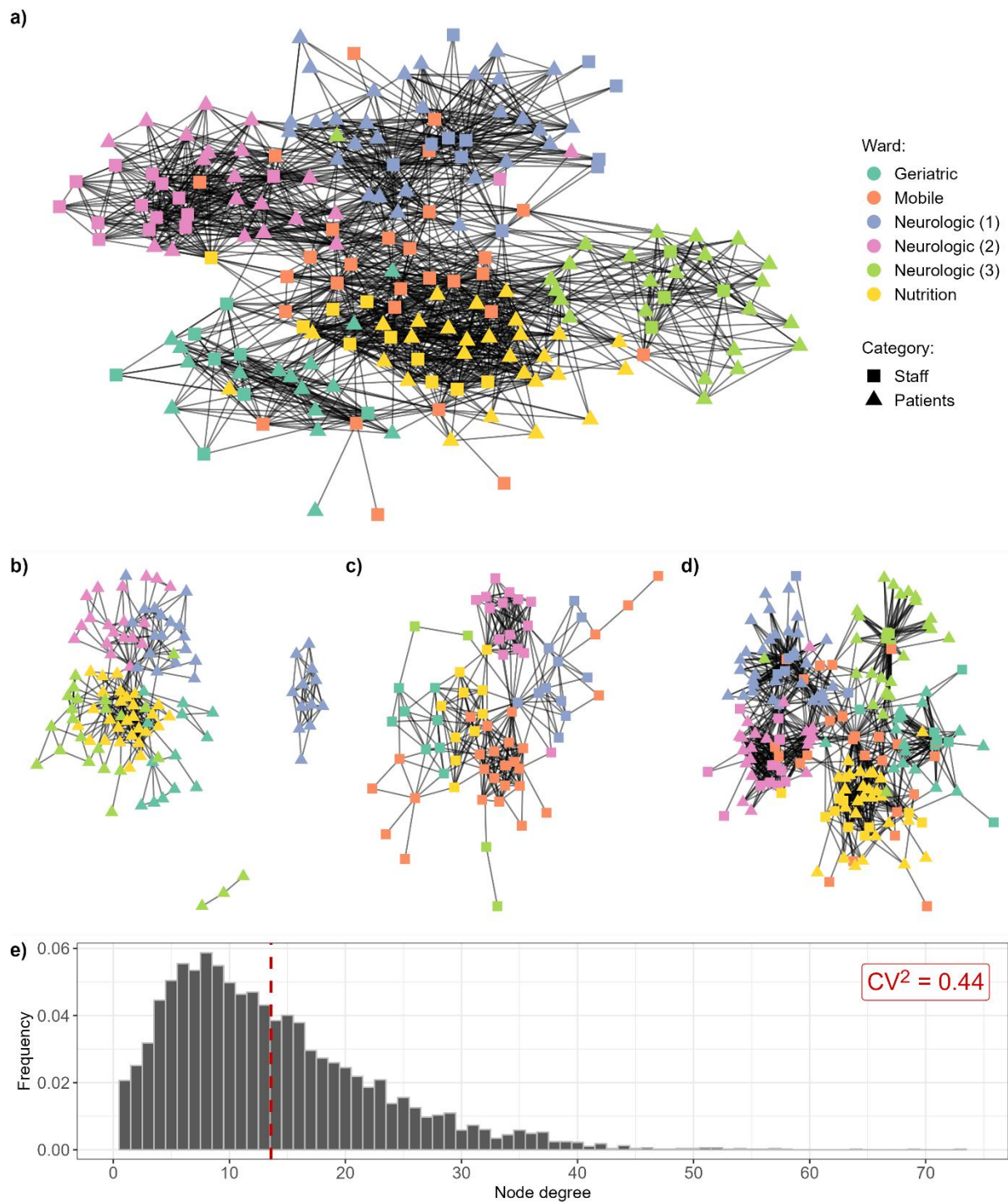
318

319 **Table 1: Summary of network characteristics for the observed i-Bird total network, patient-**
 320 **patient subgraph, staff-staff subgraph, and patient-staff subgraph.** Values were estimated
 321 for each day of the 28-days period and summarised here with the mean and standard
 322 deviation (sd). Transitivity is not calculated for the patient-staff subgraph as triangles of
 323 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)				
<i>By degree</i>	-0.13 (0.10)	0.22 (0.10)	0.14 (0.14)	-0.42 (0.14)
<i>By ward</i>	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)

324

325



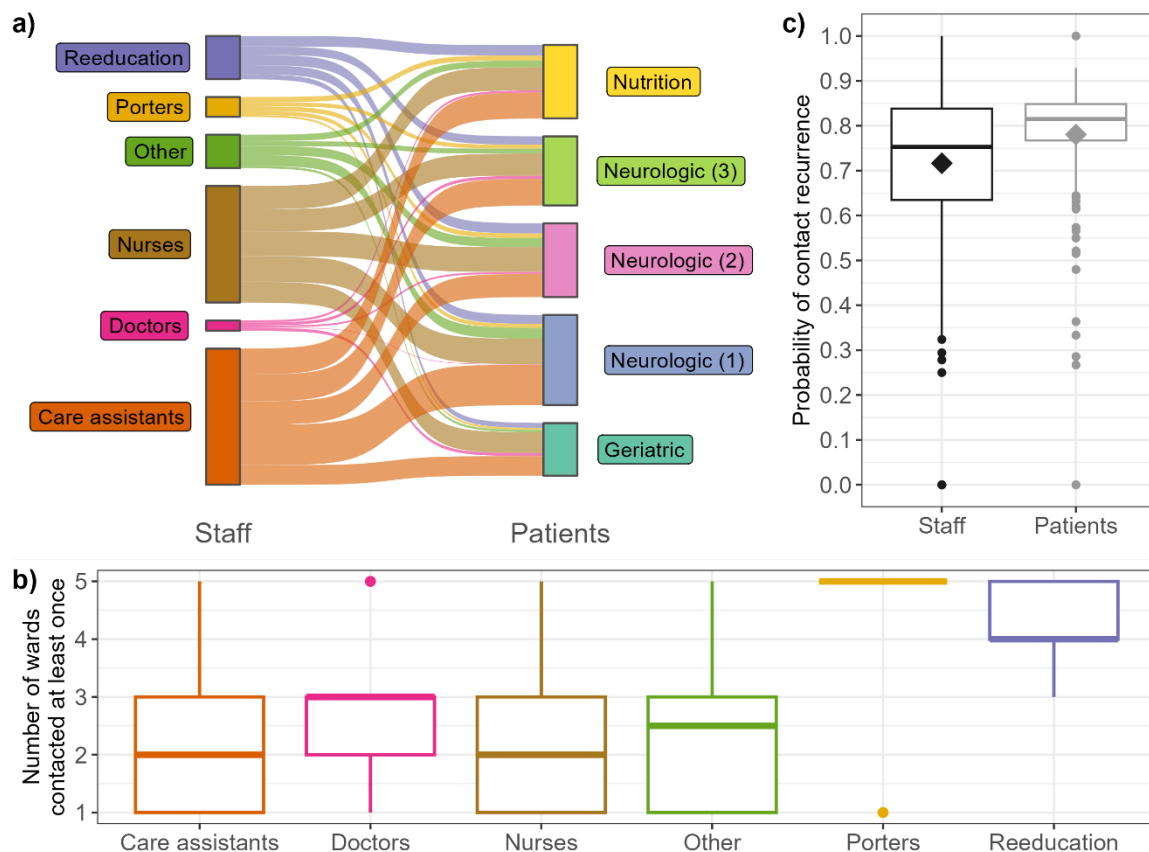
326

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

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

342
343 Overall, contacts are relatively well maintained over time, as shown by the temporal
344 correlation coefficient of 0.47 (sd: 0.11, Table 1). This corresponds to the average probability
345 that, between two subsequent days, an individual maintains the same number of unique
346 contacts, with the same individuals. This metric is highest in the patient-patient subgraph
347 (0.65, sd: 0.07) and lowest in the patient-staff subgraph (0.35, sd: 0.16), indicating that
348 patients tend to have the same contacts with each other every day, whilst contacts amongst
349 healthcare workers often vary between subsequent days. This consistency over time is
350 reflected in the high probability of recurring contacts (mean probability: 0.78 for patients, 0.71
351 for staff), although we note more variability amongst staff than patients (Figure 2c).

352
353 All the characteristics described above differ between weekdays and weekends in the network
354 and indicate that there are fewer contacts during weekends (Supplementary Table 1). This
355 difference is reflected in the temporal correlation, which tends to be high when comparing
356 Sunday to Saturday, but low when comparing Saturday to Friday and Monday to Sunday,
357 indicating that the structure of the network changes the most between these timepoints
358 (Supplementary Figure 3).



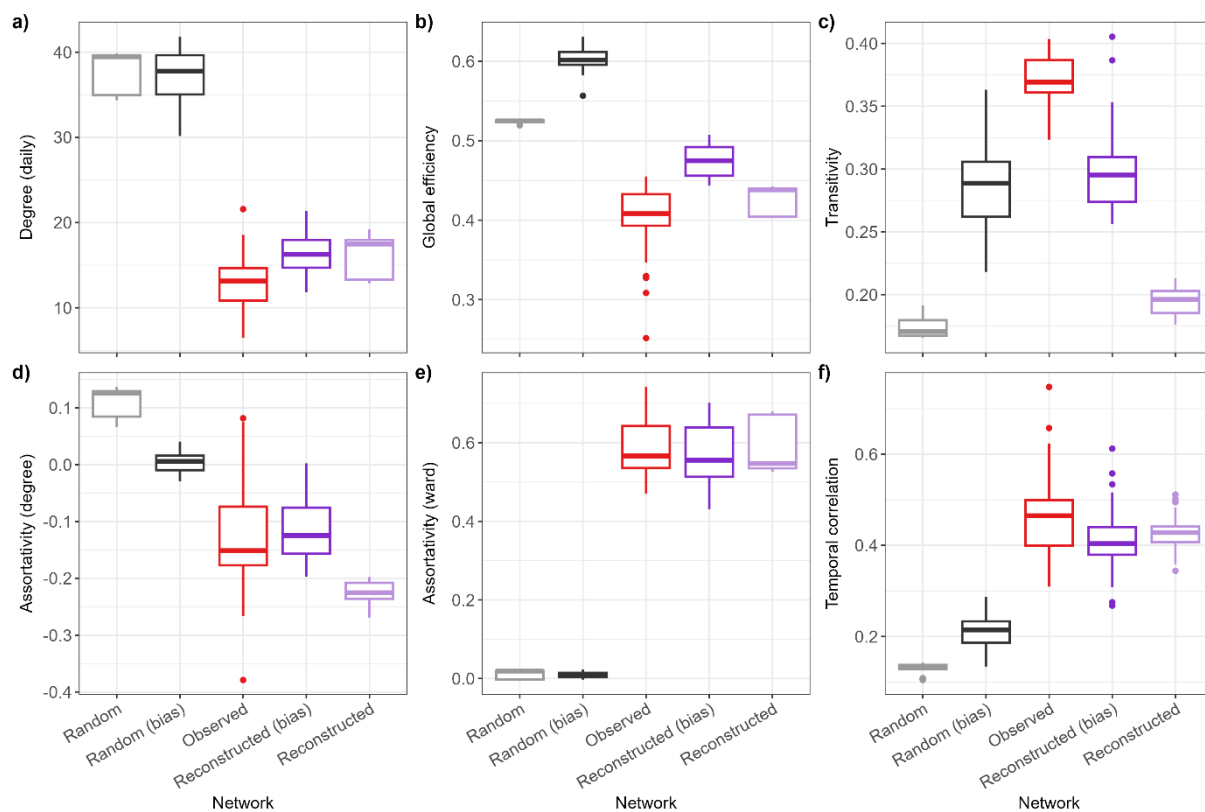
359
 360 **Figure 2: Description of contact heterogeneity and recurrence across the facility. a)**
 361 **Repartition of contacts between grouped staff professions and patient wards. A link**
 362 **between one staff category and one patient ward indicates that, at any point during the**
 363 **investigation period, a staff member from that category had a contact with a patient from that**
 364 **ward. For ease of visualisation, occupational therapists, physiotherapists, and other re-**
 365 **education staff are grouped into “Reeducation”; administrative, animation/hairdresser,**
 366 **logistic, and hospital service agents are grouped into “Other”; and nurses, head nurses, and**
 367 **students/interns are grouped into “Nurses”. Porters, doctors and care assistants are not**
 368 **grouped. b) Distribution of number of wards with which each staff member has had at least**
 369 **one contact with during the study period. c) Distribution of probabilities of recurring**
 370 **contacts.** Each observation is calculated over the entire studied period, and corresponds to
 371 the average probability for one staff or one patient to form a new contact with a previously-
 372 met individual (staff or patient) over the studied period rather than a new individual.
 373 Diamonds indicate the mean values.
 374
 375

376 **Comparison of synthetic and observed networks**

377 To illustrate and validate our algorithm, we applied it to the i-Bird network described above
378 to stochastically construct four types of synthetic networks using the estimated contact
379 parameters: 100 full reconstructed networks, 100 reconstructed networks incorporating
380 observation bias, 100 full pseudo-random networks, and 100 pseudo-random networks
381 incorporating observation bias. We expected that the characteristics of the reconstructed
382 networks with observation bias would be broadly similar to those of the observed i-Bird
383 network. Summary network characteristics are reported in Figure 3 and Supplementary Figure
384 4.

385
386 The daily degrees in the reconstructed networks were slightly higher than the observed
387 network (Figure 3a). Global efficiency was similar between the observed and reconstructed
388 networks, but slightly higher in the reconstructed network with bias (Figure 3b). This is
389 because the algorithm with bias removed individuals from the network at times when they
390 did not wear their sensor during the study, hence reducing the average distance between
391 remaining individuals. For the same reason, the density of the reconstructed network with
392 bias was slightly higher than the observed (Supplementary Figure 4). Transitivity was slightly
393 higher for the reconstructed network with observation bias than without, but lower than the
394 observed network in any case (Figure 3c), as expected since the algorithm did not take into
395 account any element of transitivity when constructing synthetic networks. Finally,
396 assortativity by degree and by ward, as well as temporal correlation, were all well preserved
397 in the reconstructed networks (Figure 3d-f). As a comparison, the random networks with or
398 without bias either substantially over- or under-estimated the values for all metrics compared
399 to the observed network (Figure 3a-f), although we note that transitivity was similar to the
400 other synthetic networks (Figure 3c).

401

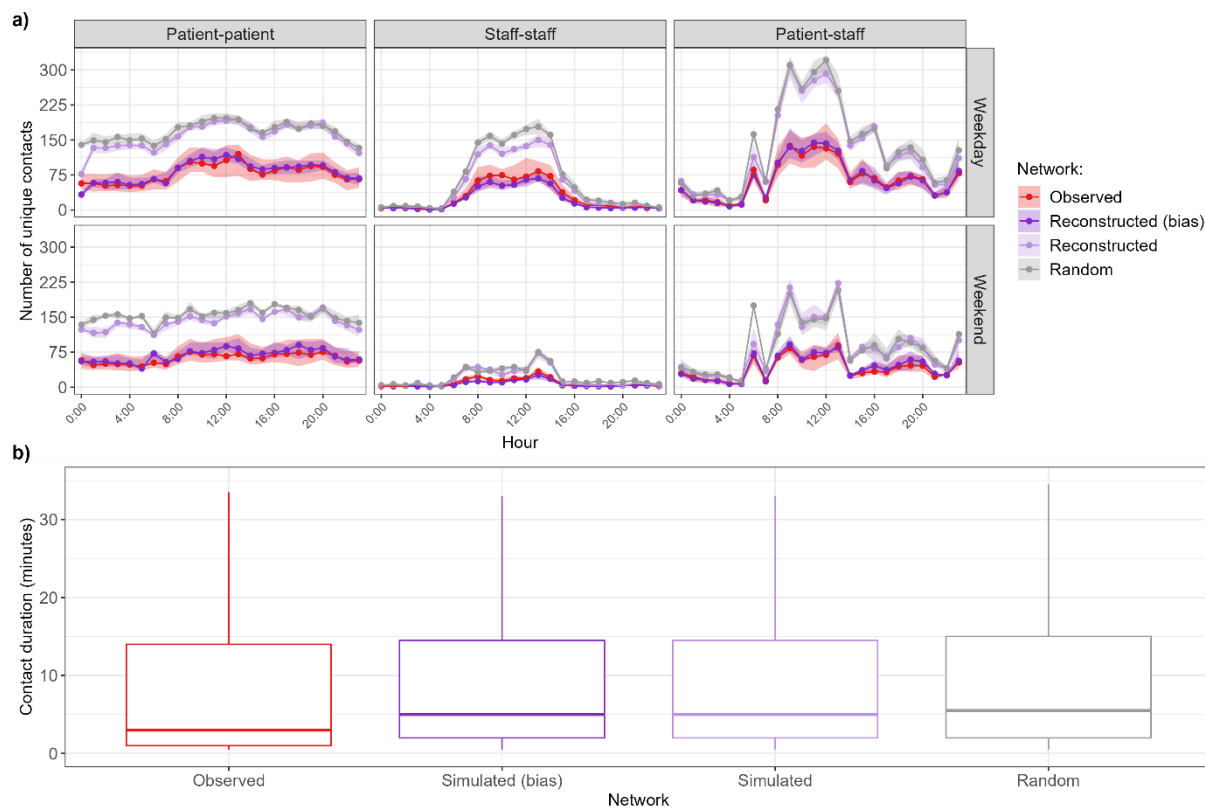


402
403 **Figure 3: Comparison of network characteristics.** The reconstructed networks with
404 observation bias exclude individuals from the network at times when they were known to not
405 wear their sensors. The random networks did not take into account the ward-level structure
406 of the contacts or the probability of recurring contacts. Boxplots for the observed network
407 show the distribution of values calculated for each day. Boxplots for all reconstructed and
408 random networks show the distribution of the median values calculated for each day across
409 100 networks.

410
411 The hourly distributions of numbers of unique patient-patient, staff-patient and staff-staff
412 contacts in the reconstructed network with bias align with those in the observed network
413 (Figure 4a). Whilst these two networks are only partially observed since individuals in the i-
414 Bird study did not have contacts recorded during all their presence days, those unobserved
415 contacts are present in the reconstructed network without bias, leading to approximately
416 twice as many contacts in that network (Figure 4a). Similarly, the random network without
417 bias which is only informed by the hourly distribution of patient-patient, staff-staff and
418 patient-staff contact rates is aligned with the reconstructed network (Figure 4a).

419

420 Finally, the distributions of contact durations in the synthetic networks were similar to the
421 distribution in the observed network, although there were slightly less contacts with short
422 durations (Figure 4b). This is because all networks sample their contact durations from a
423 lognormal distribution parameterised by the mean and variance estimated from the data,
424 which puts less emphasis on very short contacts of less than one minute (Supplementary
425 Figure 5).
426



427
428 **Figure 4: Comparison of network contact number and duration. a) Distribution of number**
429 **of unique contacts per hour, separated by type of day (weekday or weekend).** Points
430 correspond to the median, and the shaded areas correspond to the interquartile range. **b)**
431 **Distribution of contact durations.** For ease of visualisation, outliers are not shown on the
432 graph.

433
434 In supplementary analyses, we assessed the robustness of our algorithm by quantifying the
435 variability of network characteristics across 100 reconstructed networks without bias
436 (Supplementary Figure 6). The variability across reconstructed networks was not statistically
437 significant for any metric (Kruskal-Wallis test, p value > 0.05) except for assortativity by degree
438 ($p < 0.001$). We also aimed to validate our approach by generating “re-simulated” networks

439 informed by summary statistics derived from the reconstructed networks with bias. These re-
440 simulated networks are similar to the full reconstructed networks, indicating that our
441 algorithm consistently recreates realistic networks and reconstructs unobserved contacts
442 (Supplementary Figure 7). However, the number of patient-patient contacts in the re-
443 simulated networks is slightly higher than in the reconstructed networks (Supplementary
444 Figure 7).

445 Discussion

446 Summary of findings

447 In this article, we present an approach to construct stochastic synthetic temporal contact
448 networks in HCS from partially observed contact data. The i-Bird network illustrates the typical
449 complex contact structures in HCS, notably with a strong assortativity by ward, varying contact
450 rates between different staff categories and patients, and different contact structures on
451 weekends compared to weekdays. Importantly, we observed temporal correlation between
452 subsequent days in the network, and we estimated that individuals were generally more likely
453 to have contacts with other individuals they previously met rather than new individuals. Our
454 reconstruction algorithm successfully captured the heterogeneity of the observed network by
455 taking into account contact rates by hour, type of day (weekday or weekend) and staff
456 category, and probabilities of recurring contacts estimated for patients and staff. The resulting
457 reconstructed networks reproduced well the characteristics of the observed network, as well
458 as the specific distribution of unique contacts per hour.

459
460 The value of approaches to stochastically generate realistic contact networks has been
461 previously discussed for schools or workplaces [6,34], although the complexity of the contact
462 structures in those settings is arguably lower than what we observed here. While these
463 approaches extended networks by either repeating contact structures at fixed intervals or
464 randomly shuffling links [6,34], our algorithm dynamically and stochastically constructs new
465 contacts at each hour based on the empirical contact rates. Previous algorithms also
466 attempted to reconstruct missing contacts for non-participants [35]. While this was not
467 accounted for here (e.g. visitors, see below for details), here we conduct this reconstruction
468 at a higher resolution, since in reality participating individuals can also have contact data
469 missing only for some hours or days of their total presence time. Finally, the novelty of our
470 approach here is that we conduct a direct comparison between the output of our algorithm
471 and the observed contact network, as opposed to other algorithms which attempted to build
472 networks directly from contact diaries and hence did not have access to an observed network
473 for comparison [36].

474

475 **Similarities between the observed and reconstructed networks**

476 Although 90% of individuals agreed to wear a sensor during the study, the i-Bird contact
477 network was only partially observed, since the median time when contacts were not recorded
478 was 33.3% (IQR: 10.5-53.6%) of a patient's presence days (40.0%, IQR: 0-75.0% for staff). This
479 could have occurred for a number of reasons which we cannot distinguish, including depleted
480 batteries, sensor malfunction, imperfect sensor-wearing compliance, or temporary patient
481 releases from the facility (see Limitations below). However, since the average contact rates of
482 individuals did not correlate with the proportion of their presence time during which no
483 contact data were recorded (Supplementary Figure 2), it can be assumed that contact patterns
484 during unobserved times were similar to those on observed times. With that assumption, we
485 were able to reconstruct contacts at those times when individuals were present but had no
486 reported contact data. The resulting full reconstructed network is a valuable representation
487 of individual interactions, as it represents the "true" contact network, compared to the i-Bird
488 empirical network which was only partially observed. Although we were inherently limited in
489 our ability to validate this network since the real, fully observed network was not available,
490 we compared it to a re-simulated network which used the reconstructed network with
491 observation bias as input. The reconstructed and re-simulated networks without bias were
492 almost identical with regards to all the network metrics we considered (Supplementary Figure
493 7), demonstrating the consistency of our algorithm to reconstruct contacts.

494

495 The reconstructed network with bias and the observed network had similar positive
496 assortativity by ward, as expected since the input data captured the contact structure by ward.
497 The negative assortativity by degree was also similar, however we noted variability between
498 different networks generated independently by the algorithm (Supplementary Figure 6). Since
499 the algorithm did not directly account for assortativity when simulating networks, this
500 similarity stems from our use of a recurring contact probability coupled with the contact rates
501 estimated by staff categories, resulting in a non-random contact structure with regards to this
502 metric. The hourly contact distribution of patient-patient, staff-staff, and patient-staff
503 contacts was also successfully reproduced by our algorithm.

504

505 A key metric of interest here is temporal correlation, which indicates how conserved the
506 network structure is over time. This type of metric is useful to determine the efficiency of
507 disease spread across temporal networks over time [37–40]. Since our algorithm took into
508 consideration the probability of recurring contacts between individuals, our reconstructed
509 networks displayed similar temporal correlation as observed, whilst random networks
510 substantially underestimated this. This aspect is therefore an important strength of our
511 approach, compared to only using estimated average contact rates to construct synthetic
512 contacts.

513

514 **Limitations of the algorithm**

515 Density and global efficiency in the reconstructed network with bias were slightly higher than
516 in the observed network. This is a likely consequence of our observation process which forcibly
517 removed individuals from the network at times when they had no contacts recorded, hence
518 reducing the number of nodes available in the network. Simultaneously, there was still a need
519 at those times to generate some novel contacts between individuals who never previously
520 met, since the probability of recurring contacts was less than 1. Combined, these elements
521 increased the overall connectivity amongst all individuals in the reconstructed network with
522 bias. Although this could facilitate disease transmission across these reconstructed networks
523 if they are used for such purpose [41], the high assortativity by ward may counter this effect
524 by slowing down transmission across the entire healthcare facility.

525

526 Our algorithm did not specifically account for transitivity when recreating contacts. This is
527 likely why the resulting transitivity was similar to that of the random network and
528 underestimated the observed value (Figure 3). Similarly to density and global efficiency
529 mentioned above, any transitivity in the reconstructed network was likely an indirect
530 consequence of assortativity by ward, restricting the pool of available individuals to generate
531 contacts and leading to interconnectivity between individuals present in the same ward.
532 Whilst we could extend our algorithm to consider transitivity when choosing the individuals
533 to put in contact, we decided not to do this here to maximise the generalisability of our
534 approach by not requiring such highly detailed contact data. In any case, this may not
535 substantially affect disease transmission simulated across these networks, since previous

536 work has shown that transitivity is a poor predictor of the total number of individuals who
537 would be infected across the network [41].

538

539 Although our algorithm can capture individual presence and absence times, information about
540 patient temporary releases from the LTCF (e.g., for weekends with their families, or for
541 shopping outside) was not available in the i-Bird data, hence such events were not accounted
542 for here, although they may occur frequently in a LTCF. Consequently, the number of presence
543 days/hours may have been overestimated, leading to an overestimation of contact days
544 among patients. Although this is negligible when comparing the observed and reconstructed
545 network with bias, this is likely why the re-simulated networks slightly overestimated the
546 number of patient-patient contacts compared to the full reconstructed network
547 (Supplementary Figure 7). We expect that this overestimation would be absent in HCS with
548 more complete information on individual presence, or in acute care facilities with shorter
549 patient lengths of stay and where temporary releases are less common. Similarly, our
550 algorithm does not consider the contacts of visitors in the hospital, and we did not have data
551 in the i-Bird study on visitors which we would have required to validate the synthetic
552 networks. Consequently, our description of the contact structure in the LTCF is not exhaustive,
553 although this does not affect the ability of our algorithm to reproduce patient-staff contacts.

554

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

563

564 **Future work**

565 In this study, we show that our algorithm can accurately reproduce the contact structure using
566 as input contact data from a given long-term care facility. A first important next step would

567 be to repeat this analysis using data collected in a different HCS such as acute care, over a
568 different time period. This is because contact structures are known to vary between different
569 HCS such as long-term or acute, with more/less contacts between different individual
570 categories, varying recurring contact probabilities etc. Similarly, even though we tested our
571 algorithm using substantial data covering four weeks, this contact structure may not be
572 representative of other time periods. Notably, the i-Bird data we used was collected in the
573 middle of the summer, which is a holiday period in France and may have affected contact
574 patterns. Although we do not expect that our algorithm will perform differently since it has
575 been designed to be generalisable, the strengths and limitations we have highlighted above
576 may be more or less relevant in these different settings. For example, in a setting with low
577 transitivity, the fact that our algorithm underestimates this metric would be less problematic.

578

579 Here we directly re-used patient admission and discharge data as well as staff schedules to
580 identify which individuals were present in the facility at each hour, and hence whom the
581 algorithm had to build contacts for. While this choice was coherent since our aim was to
582 compare the observed and reconstructed networks, a second possible extension of our work
583 would be to simulate the presence of individuals over time. This could be implemented by
584 extracting admission and discharge rates for each category of staff and patients and using
585 these values to recreate new presence times for individuals by sampling from relevant
586 probability distributions while maintaining constraints on each population size. This would
587 allow us to further account for possible variability in the structure of the population in the
588 facility, add flexibility in building synthetic networks for settings where this data may not be
589 fully available, and hence add further stochasticity in our algorithm.

590

591 Since contact data may only be available for short periods of time (e.g. a few days [22]), a third
592 question of interest would be to understand the volume of data required to generate realistic
593 temporal contact networks using our algorithm. In our main analysis, we used the entire four
594 weeks available to both derive contact parameters and compare the reconstructed and
595 observed networks. For sensitivity, we also considered smaller time periods to calculate the
596 summary contact parameters required by the algorithm (Supplementary Figure 8). As
597 expected, this led to variability amongst the reconstructed networks depending on the length
598 of the period used, since this reduced the number of data points used to estimate the average

599 contact rates used by the algorithm. In any case, the main risk of using only a short period of
600 time is to miss out on some contacts between categories. For example, during a single week,
601 by chance there may not be any observed contact between patients from one ward w_1 and a
602 nurse from another ward w_2 , while in reality over a longer period of time we may observe a
603 few of such contacts. In that case, the algorithm will systematically assume that such contacts
604 never occur during the entire period over which the reconstructed networks are generated
605 and will therefore construct an incomplete network. A further extension of our algorithm
606 could include the possibility of creating such unobserved links, but this would still require
607 either assumptions or information on the nature of those links. Therefore, it is essential for
608 users to be confident that the data they use include contact rates for all relevant categories in
609 their setting and for typical representative days.

610
611 As discussed above, taking into account the probability for contacts to be recurring instead of
612 assuming a uniform distribution is a key element of our approach. Here, we estimate the
613 average probabilities of recurring contacts in the studied LTCF over the studied period as 0.71
614 for staff and 0.78 for patients, but we note some individual variation in this value (Figure 2,
615 interquartile range for staff: 0.63-0.84, for patients: 0.77-0.85). In addition, our estimation
616 here is made using the entire observed contact networks over the study period, but this may
617 be difficult in instances where only limited data are available. For sensitivity, we investigated
618 the impact of manually setting the probabilities to 0.1, 0.5 and 0.9 for both staff and patients
619 (Supplementary Figure 9). This led to important variations in assortativity by degree and
620 temporal correlation compared to using the estimated probability. A greater understanding
621 of this recurring contact probability in various settings would be key to better understand
622 contact formation and heterogeneity, and could be directly taken into consideration in our
623 algorithm since it has been designed to use this probability. In healthcare settings, this
624 probability could likely be estimated without requiring complete contact data, using
625 information on staff schedules and patient ward or room allocation instead.

626
627 Finally, other methodological approaches could be considered to reconstruct realistic contact
628 networks. For example, deep learning algorithms such as graph convolutional networks (GCN)
629 have become increasingly popular for this purpose, particularly in the context of infectious
630 disease transmission [42–46]. It would be interesting to compare the performance of these

631 approaches with our algorithm to estimate network characteristics and reconstruct
632 unobserved contacts. However, traditional GCN approaches do not account for temporal
633 dependencies between contacts such as the ones we observed in the i-Bird network where
634 the probability of recurring contacts plays a key role [47,48]. On the other hand, temporal
635 graph networks can capture this temporal dependency [49,50], but require substantial
636 computational resources to be applied to a network such as i-Bird, with hundreds of
637 interactions recorded every 30 seconds during several weeks. Finally, deep learning methods
638 require large amounts of training data. Democratising their use would therefore first require
639 new studies to collect close-proximity interaction data in different settings and time periods,
640 presenting further logistical challenges.

641

642 **Implications**

643 Our algorithm relies on computing summary statistics from an observed network, then using
644 these statistics to stochastically reconstruct contact networks. Such statistics can be derived
645 directly from other observed networks, as we have done here to validate our approach. In
646 that case, instead of only relying on a single observed network, our approach provides
647 multiple realistic reconstructed networks enabling to consider the impact of stochasticity of
648 the contact structure and on subsequent epidemic risk in a given setting. Our approach, by
649 providing data augmentation, also enables to infer information on potentially unobserved
650 contacts and generate extended realistic temporal dynamics over longer time periods than
651 the period of data collection.

652

653 Alternatively, summary contact statistics could be more simply collected from cross-sectional
654 surveys or even derived exclusively from individual schedules, which would not require a
655 detailed and costly follow-up using sensors. In this scenario, the only other data required
656 would be individual presence times, which should either be routinely available (e.g. in
657 healthcare settings or schools) or relatively easy to collect (e.g. in workplaces). Although as
658 mentioned in the Limitations, the amount of data our algorithm requires to generate realistic
659 networks is still unclear, our approach could ultimately be used to generate contact networks
660 from contact matrices. This would substantially facilitate research on the impact of contact
661 heterogeneity in various populations and settings, as others have previously discussed [36].

662 In conclusion, our algorithm can generate temporal contact networks in a healthcare setting
663 by taking into consideration empirically measured contact rates based on close-proximity
664 sensors, as opposed to most available packages which only construct static networks and rely
665 on hyperparameters [51,52]. These temporal networks can then be analysed with
666 mathematical models to evaluate the potential impact of interventions against disease
667 transmission [11–14]. In particular, this will improve the wider applicability of individual-based
668 model which make it possible to account for detailed contact heterogeneity in testing the
669 effect of interventions targeting highly specific individuals.

670 References

- 671 1. Keeling MJ, Rohani P. Modeling Infectious Diseases in Humans and Animals. Modeling
672 Infectious Diseases in Humans and Animals. Princeton University Press; 2008.
673 doi:10.1515/9781400841035
- 674 2. Anderson RM, May RM. Infectious diseases of humans: dynamics and control. Reprinted.
675 Oxford: Oxford Univ. Press; 2010.
- 676 3. Diekmann O, Heesterbeek JAP. Mathematical Epidemiology of Infectious Diseases: Model
677 Building, Analysis and Interpretation. John Wiley & Sons; 2000.
- 678 4. Großmann G, Backenköhler M, Wolf V. Heterogeneity matters: Contact structure and individual
679 variation shape epidemic dynamics. PLoS One. 2021;16: e0250050.
680 doi:10.1371/journal.pone.0250050
- 681 5. Machens A, Gesualdo F, Rizzo C, Tozzi AE, Barrat A, Cattuto C. An infectious disease model on
682 empirical networks of human contact: bridging the gap between dynamic network data and
683 contact matrices. BMC Infectious Diseases. 2013;13: 185. doi:10.1186/1471-2334-13-185
- 684 6. Stehlé J, Voirin N, Barrat A, Cattuto C, Colizza V, Isella L, et al. Simulation of an SEIR infectious
685 disease model on the dynamic contact network of conference attendees. BMC Medicine.
686 2011;9: 87. doi:10.1186/1741-7015-9-87
- 687 7. Kiss IZ, Miller JC, Simon PL, others. Mathematics of epidemics on networks. Cham: Springer.
688 2017;598: 31.
- 689 8. Danon L, Ford AP, House T, Jewell CP, Keeling MJ, Roberts GO, et al. Networks and the
690 Epidemiology of Infectious Disease. Interdiscip Perspect Infect Dis. 2011;2011: 284909.
691 doi:10.1155/2011/284909
- 692 9. Mossong J, Hens N, Jit M, Beutels P, Auranen K, Mikolajczyk R, et al. Social Contacts and Mixing
693 Patterns Relevant to the Spread of Infectious Diseases. PLOS Medicine. 2008;5: e74.
694 doi:10.1371/journal.pmed.0050074
- 695 10. Keeling MJ, Eames KTD. Networks and epidemic models. Journal of The Royal Society Interface.
696 2005;2: 295–307. doi:10.1098/rsif.2005.0051
- 697 11. Bansal S, Read J, Pourbohloul B, Meyers LA. The dynamic nature of contact networks in
698 infectious disease epidemiology. Journal of Biological Dynamics. 2010;4: 478–489.
699 doi:10.1080/17513758.2010.503376
- 700 12. Masuda N, Holme P. Predicting and controlling infectious disease epidemics using temporal
701 networks. F1000Prime Rep. 2013;5: 6. doi:10.12703/P5-6
- 702 13. Gross T, D’Lima CJD, Blasius B. Epidemic Dynamics on an Adaptive Network. Phys Rev Lett.
703 2006;96: 208701. doi:10.1103/PhysRevLett.96.208701
- 704 14. Valdano E, Poletto C, Giovannini A, Palma D, Savini L, Colizza V. Predicting Epidemic Risk from
705 Past Temporal Contact Data. PLOS Computational Biology. 2015;11: e1004152.
706 doi:10.1371/journal.pcbi.1004152

- 707 15. Holme P, Saramäki J. Temporal networks. *Physics Reports*. 2012;519: 97–125.
708 doi:10.1016/j.physrep.2012.03.001
- 709 16. Hornbeck T, Naylor D, Segre AM, Thomas G, Herman T, Polgreen PM. Using Sensor Networks to
710 Study the Effect of Peripatetic Healthcare Workers on the Spread of Hospital-Associated
711 Infections. *Journal of Infectious Diseases*. 2012;206: 1549–1557. doi:10.1093/infdis/jis542
- 712 17. Obadia T, Silhol R, Opatowski L, Temime L, Legrand J, Thiébaud ACM, et al. Detailed Contact
713 Data and the Dissemination of *Staphylococcus aureus* in Hospitals. Salathé M, editor. *PLoS*
714 *Comput Biol*. 2015;11: e1004170. doi:10.1371/journal.pcbi.1004170
- 715 18. Salathé M, Kazandjieva M, Lee JW, Levis P, Feldman MW, Jones JH. A high-resolution human
716 contact network for infectious disease transmission. *Proceedings of the National Academy of*
717 *Sciences*. 2010;107: 22020–22025. doi:10.1073/pnas.1009094108
- 718 19. Min-Allah N, Alahmed BA, Albreek EM, Alghamdi LS, Alawad DA, Alharbi AS, et al. A survey of
719 COVID-19 contact-tracing apps. *Computers in Biology and Medicine*. 2021;137: 104787.
720 doi:10.1016/j.combiomed.2021.104787
- 721 20. Eames K, Bansal S, Frost S, Riley S. Six challenges in measuring contact networks for use in
722 modelling. *Epidemics*. 2015;10: 72–77. doi:10.1016/j.epidem.2014.08.006
- 723 21. Read JM, Edmunds WJ, Riley S, Lessler J, Cummings D a. T. Close encounters of the infectious
724 kind: methods to measure social mixing behaviour. *Epidemiology & Infection*. 2012;140: 2117–
725 2130. doi:10.1017/S0950268812000842
- 726 22. Vanhems P, Barrat A, Cattuto C, Pinton J-F, Khanafer N, Régis C, et al. Estimating Potential
727 Infection Transmission Routes in Hospital Wards Using Wearable Proximity Sensors. Viboud C,
728 editor. *PLoS ONE*. 2013;8: e73970. doi:10.1371/journal.pone.0073970
- 729 23. Stehlé J, Voirin N, Barrat A, Cattuto C, Isella L, Pinton J-F, et al. High-Resolution Measurements
730 of Face-to-Face Contact Patterns in a Primary School. *PLOS ONE*. 2011;6: e23176.
731 doi:10.1371/journal.pone.0023176
- 732 24. Smieszek T, Castell S, Barrat A, Cattuto C, White PJ, Krause G. Contact diaries versus wearable
733 proximity sensors in measuring contact patterns at a conference: method comparison and
734 participants' attitudes. *BMC Infectious Diseases*. 2016;16: 341. doi:10.1186/s12879-016-1676-y
- 735 25. Watts DJ, Strogatz SH. Collective dynamics of 'small-world' networks. *Nature*. 1998;393: 440–
736 442. doi:10.1038/30918
- 737 26. Almutiry W, Deardon R. Contact network uncertainty in individual level models of infectious
738 disease transmission. *Stat Commun Infect Dis*. 2021;13: 20190012. doi:10.1515/scid-2019-0012
- 739 27. Shirley MDF, Rushton SP. The impacts of network topology on disease spread. *Ecological*
740 *Complexity*. 2005;2: 287–299. doi:10.1016/j.ecocom.2005.04.005
- 741 28. Gimma A, Munday JD, Wong KLM, Coletti P, van Zandvoort K, Prem K, et al. Changes in social
742 contacts in England during the COVID-19 pandemic between March 2020 and March 2021 as
743 measured by the CoMix survey: A repeated cross-sectional study. *PLoS Med*. 2022;19:
744 e1003907. doi:10.1371/journal.pmed.1003907

- 745 29. Mousa A, Winskill P, Watson OJ, Ratmann O, Monod M, Ajelli M, et al. Social contact patterns
746 and implications for infectious disease transmission – a systematic review and meta-analysis of
747 contact surveys. Rodriguez-Barraquer I, Serwadda DM, editors. eLife. 2021;10: e70294.
748 doi:10.7554/eLife.70294
- 749 30. Hoang T, Coletti P, Melegaro A, Wallinga J, Grijalva CG, Edmunds JW, et al. A Systematic Review
750 of Social Contact Surveys to Inform Transmission Models of Close-contact Infections.
751 Epidemiology. 2019;30: 723–736. doi:10.1097/EDE.0000000000001047
- 752 31. Duval A, Obadia T, Martinet L, Boëlle P-Y, Fleury E, Guillemot D, et al. Measuring dynamic social
753 contacts in a rehabilitation hospital: effect of wards, patient and staff characteristics. Scientific
754 Reports. 2018;8: 1686. doi:10.1038/s41598-018-20008-w
- 755 32. R Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria: R
756 Foundation for Statistical Computing; 2022. Available: <https://www.R-project.org/>
- 757 33. Csardi G, Nepusz T. The igraph software package for complex network research. InterJournal.
758 2006;Complex Systems: 1695.
- 759 34. Colosi E, Bassignana G, Contreras DA, Poirier C, Boëlle P-Y, Cauchemez S, et al. Screening and
760 vaccination against COVID-19 to minimise school closure: a modelling study. The Lancet
761 Infectious Diseases. 2022;22: 977–989. doi:10.1016/S1473-3099(22)00138-4
- 762 35. Génois M, Vestergaard CL, Cattuto C, Barrat A. Compensating for population sampling in
763 simulations of epidemic spread on temporal contact networks. Nat Commun. 2015;6: 8860.
764 doi:10.1038/ncomms9860
- 765 36. Mastrandrea R, Barrat A. How to Estimate Epidemic Risk from Incomplete Contact Diaries Data?
766 PLOS Computational Biology. 2016;12: e1005002. doi:10.1371/journal.pcbi.1005002
- 767 37. Tang J, Scellato S, Musolesi M, Mascolo C, Latora V. Small-world behavior in time-varying
768 graphs. Phys Rev E. 2010;81: 055101. doi:10.1103/PhysRevE.81.055101
- 769 38. Kretzschmar M, Morris M. Measures of concurrency in networks and the spread of infectious
770 disease. Mathematical Biosciences. 1996;133: 165–195. doi:10.1016/0025-5564(95)00093-3
- 771 39. Read JM, Eames KTD, Edmunds WJ. Dynamic social networks and the implications for the
772 spread of infectious disease. J R Soc Interface. 2008;5: 1001–1007. doi:10.1098/rsif.2008.0013
- 773 40. Smieszek T, Fiebig L, Scholz RW. Models of epidemics: when contact repetition and clustering
774 should be included. Theor Biol Med Model. 2009;6: 11. doi:10.1186/1742-4682-6-11
- 775 41. Pérez-Ortiz M, Manescu P, Caccioli F, Fernández-Reyes D, Nachev P, Shawe-Taylor J. Network
776 topological determinants of pathogen spread. Sci Rep. 2022;12: 7692. doi:10.1038/s41598-022-
777 11786-5
- 778 42. Fritz C, Dorigatti E, Rügamer D. Combining graph neural networks and spatio-temporal disease
779 models to improve the prediction of weekly COVID-19 cases in Germany. Sci Rep. 2022;12:
780 3930. doi:10.1038/s41598-022-07757-5
- 781 43. Gao J, Sharma R, Qian C, Glass LM, Spaeder J, Romberg J, et al. STAN: spatio-temporal attention
782 network for pandemic prediction using real-world evidence. Journal of the American Medical
783 Informatics Association. 2021;28: 733–743. doi:10.1093/jamia/ocaa322

- 784 44. Panagopoulos G, Nikolentzos G, Vazirgiannis M. Transfer Graph Neural Networks for Pandemic
785 Forecasting. Proceedings of the AAAI Conference on Artificial Intelligence. 2021;35: 4838–4845.
786 doi:10.1609/aaai.v35i6.16616
- 787 45. Kapoor A, Ben X, Liu L, Perozzi B, Barnes M, Blais M, et al. Examining COVID-19 Forecasting
788 using Spatio-Temporal Graph Neural Networks. arXiv; 2020. doi:10.48550/arXiv.2007.03113
- 789 46. Zhao G, Jia P, Zhou A, Zhang B. InfGCN: Identifying influential nodes in complex networks with
790 graph convolutional networks. Neurocomputing. 2020;414: 18–26.
791 doi:10.1016/j.neucom.2020.07.028
- 792 47. Li L, Zhou J, Jiang Y, Huang B. Propagation source identification of infectious diseases with
793 graph convolutional networks. Journal of Biomedical Informatics. 2021;116: 103720.
794 doi:10.1016/j.jbi.2021.103720
- 795 48. Ni Q, Wu X, Chen H, Jin R, Wang H. Spatial-temporal deep learning model based rumor source
796 identification in social networks. J Comb Optim. 2023;45: 86. doi:10.1007/s10878-023-01018-5
- 797 49. Holme P. Modern temporal network theory: a colloquium. Eur Phys J B. 2015;88: 234.
798 doi:10.1140/epjb/e2015-60657-4
- 799 50. Tang J, Leontiadis I, Scellato S, Nicosia V, Mascolo C, Musolesi M, et al. Applications of
800 Temporal Graph Metrics to Real-World Networks. In: Holme P, Saramäki J, editors. Temporal
801 Networks. Berlin, Heidelberg: Springer; 2013. pp. 135–159. doi:10.1007/978-3-642-36461-7_7
- 802 51. Prettejohn B, Berryman M, McDonnell M. Methods for Generating Complex Networks with
803 Selected Structural Properties for Simulations: A Review and Tutorial for Neuroscientists.
804 Frontiers in Computational Neuroscience. 2011;5. Available:
805 <https://www.frontiersin.org/articles/10.3389/fncom.2011.00011>
- 806 52. Hagberg AA, Schult DA, Swart PJ. Exploring Network Structure, Dynamics, and Function using
807 NetworkX. In: Varoquaux G, Vaught T, Millman J, editors. Proceedings of the 7th Python in
808 Science Conference. Pasadena, CA USA; 2008. pp. 11–15.