Génois and Barrat *EPJ Data Science* (2018) 7:11
https://doi.org/10.1140/epjds/s13688-018-0140-1

EPJ Data Science
a SpringerOpen Journal

REGULAR ARTICLE Open Access

# Can co-location be used as a proxy for face-to-face contacts?

Mathieu Génois[1,2]* and Alain Barrat[1,3]

*Correspondence:
mathieu.genois@gesis.org
[1] CNRS, CPT, Aix Marseille Univ, Université de Toulon, Marseille, France
[2] GESIS, Leibniz Institute for the Social Sciences, Köln, Germany
Full list of author information is available at the end of the article

**Abstract**

Technological advances have led to a strong increase in the number of data collection efforts aimed at measuring co-presence of individuals at different spatial resolutions. It is however unclear how much co-presence data can inform us on actual face-to-face contacts, of particular interest to study the structure of a population in social groups or for use in data-driven models of information or epidemic spreading processes. Here, we address this issue by leveraging data sets containing high resolution face-to-face contacts as well as a coarser spatial localisation of individuals, both temporally resolved, in various contexts. The co-presence and the face-to-face contact temporal networks share a number of structural and statistical features, but the former is (by definition) much denser than the latter. We thus consider several down-sampling methods that generate surrogate contact networks from the co-presence signal and compare them with the real face-to-face data. We show that these surrogate networks reproduce some features of the real data but are only partially able to identify the most central nodes of the face-to-face network. We then address the issue of using such down-sampled co-presence data in data-driven simulations of epidemic processes, and in identifying efficient containment strategies. We show that the performance of the various sampling methods strongly varies depending on context. We discuss the consequences of our results with respect to data collection strategies and methodologies.

**Keywords:** Face-to-face contacts; Co-presence; Digital epidemiology; Complex networks

## 1 Introduction

In the recent years, several methods have been developed to gather quantitative data on human interactions using wearable sensors and complement more traditional methods based on surveys [1–3]. Current data collection methods range from the use of Bluetooth or WiFi signals in mobile phones [4–9] to the specific development of dedicated sociometric sensors [10–19] and enable researchers to record and measure physical proximity events between individuals in various social contexts. Depending on the specific technology considered however, spatial resolution varies and the resulting "contacts" detected can range from co-presence in a room or a part of a building to close face-to-face encounters. The resulting data is often temporally resolved and has been increasingly used in various contexts including the study of human behaviour, the validation of models of human interactions and data-driven models of epidemic spreading [3, 20, 21].

Despite the increasing availability of techniques to measure even high-resolution temporal contact networks however, a number of limitations remain. In particular, measures cannot be carried out for arbitrarily large population sizes. It is thus of crucial interest to infer contacts or build contact proxies from data with lower spatial resolution data or coming from other sources. In this spirit, several studies have considered the issue of inferring social ties from email exchanges [22], mobile phone data [23], or co-location at geographic scale [24]. Other works try to infer close proximity in specific settings from individual attributes [25] or from a very precise localisation of individuals [17], or, at geographical scale, from the similarity of the WiFi signals received from a large enough number of WiFi routers [26].

Here instead, we do not try to infer specific contacts between pairs of individuals but rather investigate if a coarse co-location information on individuals allows us to reach an overall picture of the contact patterns in the population of interest. Since gathering large-scale data about localisation is much easier than recording face-to-face contacts, a method to infer general characteristics of the latter from the former would enable faster, larger and more diverse data collections about human behaviour. To this aim, we leverage several data sets collected by the SocioPatterns collaboration [13, 27] in various contexts: these data include both detailed information about close, face-to-face encounters between individuals and a location tracking of individuals with low spatial resolution. It is thus possible to build two temporal networks where nodes represent individuals and links correspond respectively to a face-to-face contact or to a co-presence event, where co-presence is defined with respect to the localisation of two individuals within the same spatial area. We first compare the structural and statistical properties of these two temporal networks and show that they share some important properties, although the co-presence network is much denser, due to the lower spatial resolution involved in its definition. We thus investigate several methods of down-sampling the co-presence signal in order to create surrogate contact networks, in the spirit of [28, 29], and compare these surrogate data to the actual networks of face-to-face contacts. We focus first on several statistical characteristics of temporal and aggregated networks, and quantify the ability to identify central nodes in the contact network from the surrogate data. We then consider the possibility to use the surrogate data in numerical simulations of data-driven models for epidemic spread. In particular, we compare the outcome of simulations of a standard model of epidemic propagation when surrogate or actual contact data are used, and we explore the possibility to identify the most efficient containment strategies from this limited information [30]. Our results turn out to depend strongly on the data collection context, highlighting the limitations of coarse co-presence networks with respect to detailed face-to-face data.

## 2 The co-presence network

### 2.1 Data sets

We use data collected by the SocioPatterns collaboration in various contexts. T hese data were gathered using wearable sensors able to detect face-to-face close range proximity (1.5 m) of participants wearing the sensors on their chests. In addition, the sensors broadcast a signal that can be received by RFID readers located in the environment. In open space, each reader can receive signals from sensors situated within a range of $\sim$30 m, while the actual reception range in a building depends on its specific structure and on the nature of its walls, floors and ceilings. Each reader thus defines a coarse spatial area and

**Table 1** Characteristics of the data sets

| Data set | Location | Year | $N_p$ | $N_a$ | $T$ | Ref |
|---|---|---|---|---|---|---|
| InVS13 | Fr. Health Obs. | 2013 | 92 | 27 | 2 weeks | [31] |
| InVS15 | Fr. Health Obs. | 2015 | 232 | 45 | 2 weeks | |
| LH10 | Hospital | 2010 | 81 | 8 | 3 days | [32] |
| LyonSchool | Primary school | 2009 | 242 | 15 | 2 days | [33] |
| SFHH | Conference | 2009 | 403 | 12 | 2 days | [34] |
| Thiers13 | High school | 2013 | 326 | 18 | 1 week | [35] |

$N_p$ is the number of participants, $N_a$ the number of RFID readers, $T$ the total duration of the data collection.

the sensors' signals can be followed when the individuals carrying them change area. For each sensor, we define its "spatial location" at each time as the set of readers receiving its broadcasted signal at this time, and we define two individuals to be in co-presence if they share the same spatial location, i.e., the same exact set of readers have received signals from both individuals.

We use data sets from various social contexts: a workplace, with data collected in two different years (InVS13, InVS15), a hospital (LH10), a primary school (LyonSchool), a scientific conference (SFHH) and a high school (Thiers13), see Table 1. In each case, we thus consider a temporal network of face-to-face contacts and a temporal network of co-presence between individuals, both at the temporal resolution of 20 s. A contact (resp. co-presence) event between two individuals is then defined as a set of successive time-windows of 20 s during which the individuals are detected in contact (resp. co-presence), while they are not in the preceding nor in the next 20 s time window. A contact or co-presence event therefore has a certain duration that is a multiple of 20 s, and can be formally represented as the quadruplet $(i, j, t, \tau)$, for a contact occurring between nodes $i$ and $j$, starting at time $t$ and with a duration $\tau$. While the conference data does not include any other information on the participants and does not exhibit any particular group structure [36], the other populations under study can each be divided into groups: departments for the workplace, classes for the school and the high school, and roles (patients, doctors, nurses) in the hospital. In these cases, the overall structure of networks aggregated over a certain time window can be summarised, in addition to usual quantities such as the density, the clustering coefficient or the degree distribution, by contact (resp. co-presence) matrices that give the fraction of pairs of individuals of different groups who have been in contact (resp. in co-presence). Moreover, temporal features of interest include the distributions of durations of contact or co-presence events, of the time elapsed between successive events, of the numbers and aggregated durations of such events between pairs of individuals (the latter quantity yields a natural definition of the weight of a link between individuals in the aggregated network).

We will show in the main text the results corresponding to the InVS15 data set, and we refer to the Additional file 1 for the results obtained with the other data sets. We make also available as Additional file 1 the temporally resolved contact and co-presence networks.

## 2.2 Co-presence and contact networks

We first compare some features of the co-presence and contact networks, both temporal and for networks aggregated either on the whole data gathering period or over daily temporal windows. We show in Fig. 1 the distributions of event and inter-event duration, as well as the distributions of number and cumulative duration of events for individual

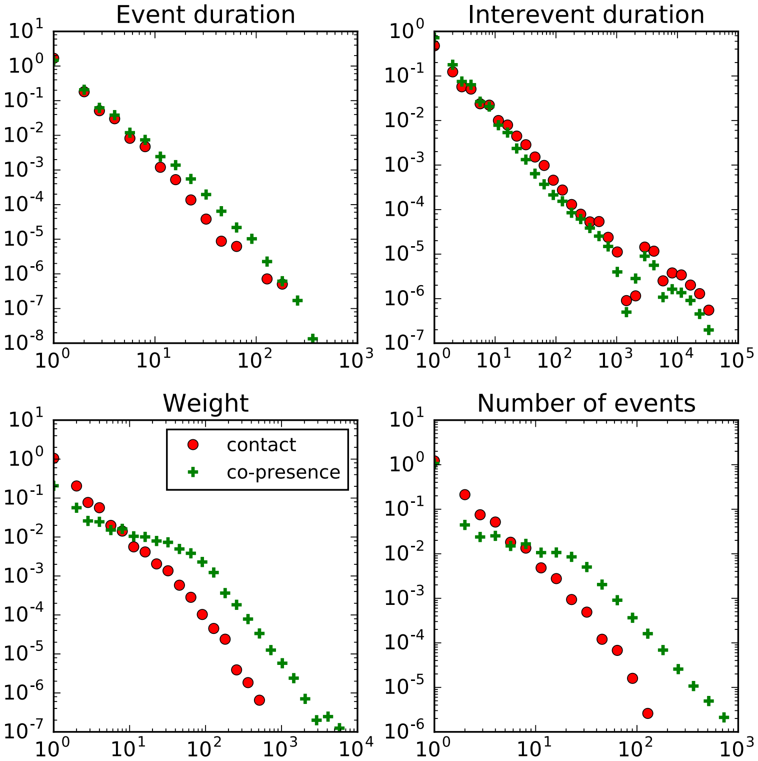

**Figure 1** Temporal distributions for co-presence and contact events—InVS15. We show for both the contact and co-presence of the same data set the distributions of event and inter-event duration, link weights (as total contact duration) and number of contacts per link.

pairs. The co-presence events show broad distributions of these quantities, similarly to the contact events and with similar slopes: using only co-presence data yields approximate information on the functional shape of the contact duration distributions. The distributions of durations and numbers of events are however typically broader for co-presence, with heavier tails, and the distribution of inter-event durations tend to be less broad (see also Additional file 1). This is not surprising as the criterion for being in co-presence is less strict than for being in contact. We observe the strongest differences between co-presence and contact distribution functional shapes for the primary school data. This could be related by the fact that the spatial resolution is in that case quite low, with all the schoolyard being covered by one single reader, and some readers covering more than one classroom. Overall, using only co-presence data would lead to over-estimations of the contact durations and aggregate durations.

We compare moreover in Figs. 2–3 and Tables 2–3 the overall structures of the contact and co-presence networks, aggregated over daily time windows. The co-presence aggregated networks are much denser than the contact network, with a larger average degree, a larger average clustering coefficient and larger cliques, as expected once again given the lower spatial resolution required for co-presence events. In some cases (school, conference), the aggregated networks are even close to being fully connected (see for illustration Fig. 3). Despite this strong difference in the overall density of links, the contact and co-presence matrices giving the density of links between and within each group, averaged across days, are very similar (Table 2). The similarity is particularly high for the hospital

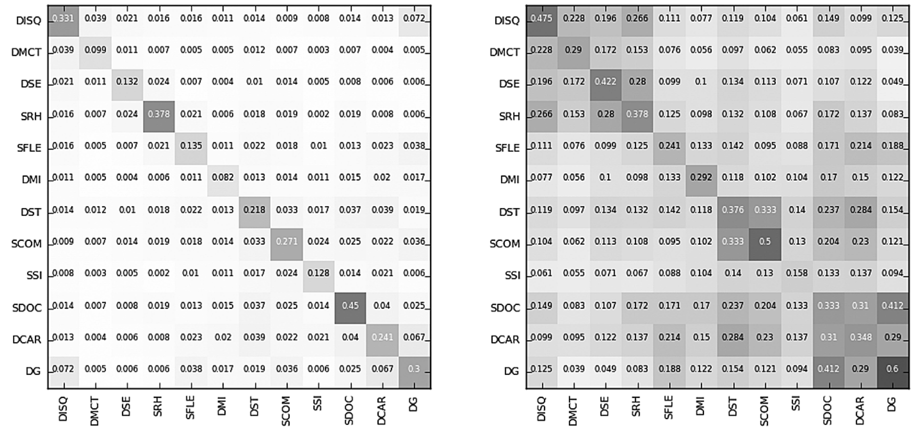

**Figure 2** Contact and co-presence matrices—InVS15. Comparison between the average matrices of link density for the contacts and the co-presence daily aggregated networks. Values are averaged over all days of the data collection. Both plots have the same colour scale.

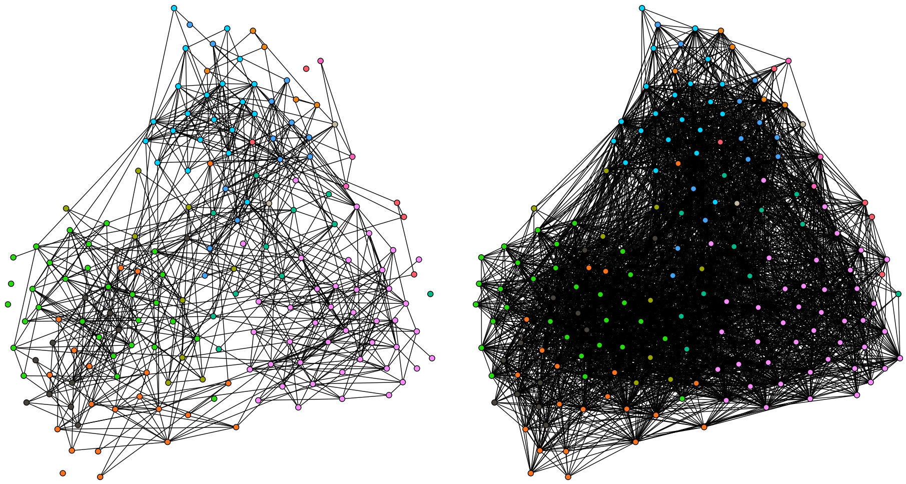

**Figure 3** Day 2 contact and co-presence networks—InVS15. This figure highlights the difference in terms of link density when comparing contact and co-presence daily networks. Different node colours correspond to the different administrative departments.

data and, even for the lower value obtained for the high school data, the matrices displayed in the Additional file 1 show that the overall structure in classes and groups of classes can be inferred from the co-presence data alone.

Given the simultaneous discrepancies in density values and similarities in the networks group structures, we investigate if the data exhibits a scaling law between the number of individuals present in an area and their contact activity, as found at geographical scale in phone communication [37] and Twitter data [38]. Figure 4 and the similar figures shown in Additional file 1 show the results obtained in the various contexts. Apart from the office cases (InVS13 and InVS15), we observe indeed a correlation between the median of the number of contacts and the number of individuals present. This correlation exhibits a power law shape, with an exponent around 1.5 (see figures in Additional file 1). However, huge, context-dependent fluctuations are observed. For instance, in the InVS15 case,

**Table 2** Similarity between contact matrices

| | InVS13 | InVS15 | LH10 | LyonSchool | Thiers13 |
|---|---|---|---|---|---|
| Co-presence | 0.790 | 0.710 | 0.968 | 0.706 | 0.681 |
| Sampling 1 | 0.946 | 0.829 | 0.960 | 0.845 | 0.857 |
| Sampling 2 | 0.958 | 0.901 | 0.894 | 0.945 | 0.937 |
| Sampling 3 | 0.888 | 0.816 | 0.958 | 0.738 | 0.691 |

For each data set we compute the cosine similarity between the average daily contact matrix and the co-presence matrix, as well as for the contact matrices obtained for each sampling method of the co-presence data, averaged over 100 realisations for each sampling method. To compute the cosine similarities, each matrix is first transformed into a vector by concatenating its rows.

**Table 3** Characteristics of the contact, co-presence, and sampled co-presence networks

| | InVS13 | InVS15 | LH10 | LyonSchool | SFHH | Thiers13 |
|---|---|---|---|---|---|---|
| $\bar{k}_c$ | 2.9 | 6.4 | 14.0 | 47.3 | 28.8 | 13.5 |
| $\bar{k}_\ell$ | 20.9 | 35.0 | 18.2 | 194.5 | 234.3 | 126.8 |
| $\bar{k}_1$ | 5.8 | 14.2 | 14.4 | 101.3 | 116.7 | 52.2 |
| $\bar{k}_2$ | 0.9 | 3.6 | 7.7 | 36.9 | 45.2 | 5.2 |
| $\bar{k}_3$ | 5.3 | 5.0 | 14.0 | 21.2 | 40.4 | 4.1 |
| $\rho_c$ | 0.030 | 0.028 | 0.175 | 0.196 | 0.072 | 0.041 |
| $\rho_\ell$ | 0.211 | 0.152 | 0.227 | 0.807 | 0.807 | 0.383 |
| $\rho_1$ | 0.058 | 0.061 | 0.179 | 0.420 | 0.290 | 0.158 |
| $\rho_2$ | 0.009 | 0.016 | 0.097 | 0.153 | 0.112 | 0.016 |
| $\rho_3$ | 0.054 | 0.022 | 0.175 | 0.088 | 0.101 | 0.013 |
| $\omega_c$ | 4.4 | 7.6 | 14.3 | 22.5 | 11.0 | 9.4 |
| $\omega_\ell$ | 18.8 | 38.7 | 22.7 | 141.0 | * | 74.6 |
| $\omega_1$ | 6.6 | 10.3 | 17.2 | 41.5 | 34.7 | 33.8 |
| $\omega_2$ | 3.0 | 5.3 | 8.6 | 12.8 | 12.2 | 3.9 |
| $\omega_3$ | 5.5 | 4.8 | 17.1 | 6.3 | 9.0 | 3.8 |
| $\bar{c}_c$ | 0.178 | 0.239 | 0.428 | 0.520 | 0.260 | 0.379 |
| $\bar{c}_\ell$ | 0.417 | 0.409 | 0.491 | 0.868 | 0.880 | 0.581 |
| $\bar{c}_1$ | 0.255 | 0.266 | 0.432 | 0.596 | 0.442 | 0.586 |
| $\bar{c}_2$ | 0.045 | 0.139 | 0.309 | 0.370 | 0.212 | 0.092 |
| $\bar{c}_3$ | 0.205 | 0.101 | 0.426 | 0.193 | 0.161 | 0.047 |

We compare the average degree ($\bar{k}$) network density ($\rho$), clique number ($\omega$) and average clustering ($\bar{c}$) of daily aggregated networks, for the contact network ($c$ subscript), the co-presence network ($\ell$ subscript), and the sampled co-presence networks (subscripts 1 to 3 according to the sampling method). Values are averaged over all the days of the study. In the case of SFHH, since on the second day there was activity only during the morning, only the values of the first day are reported. *The network is too large and too dense for the clique number to be determined in reasonable time via the usual algorithm.

the trend is strongly influenced by the numerous instances of an absence of contacts despite potentially large values of the number of individuals present in the area. This is a consequence of the fact that a given reader can receive signals from the sensors of individuals located in different offices. In other areas such as a cafeteria, many more contacts occur with potentially a similar or even smaller number of individuals. Overall, very large fluctuations of the number of contacts, at given number of individuals present, are thus observed, because on the one hand of the low spatial resolution of the co-presence data, and on the other hand of the variety of contexts corresponding to the areas covered by different RFID readers. The stronger correlation is observed for the SFHH conference data, probably because the various areas covered by the readers corresponded to similar contexts, namely different areas of the exhibition and poster rooms.

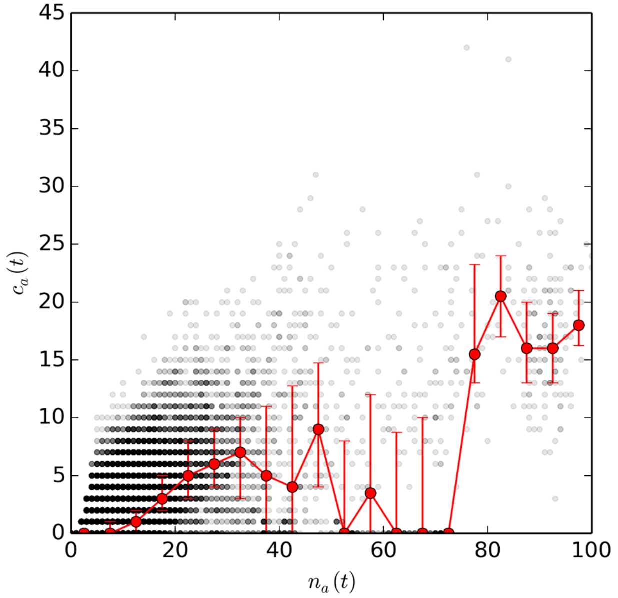

**Figure 4** Number of contacts as a function of the number of individuals present—InVS15. We plot the number of contacts $c_a(t)$ occurring at time $t$ in a certain area $a$ as a function of the number of individuals $n_a(t)$ present at the same time in $a$. The red line shows the median of the scatter plot, with error bars defined by the 25% and 75% percentiles.

## 3 Sampling co-presence data

### 3.1 Sampling methods

As the temporal network of co-presence bears some similarities with the actual contact data, but contains much more events and leads to much denser aggregated networks, we consider the possibility to down-sample the co-presence data: for each pair of individuals, each contact event is indeed included in a co-presence event of the same individuals. Each co-presence event might thus correspond to one or more contact events. As we cannot determine exactly the correct down-sampling to be performed if we have access only to co-presence data, we study here three simple sampling methods. We remind here that we do not try to infer the real contacts but rather to obtain a down-sampled version of the co-presence network that is statistically similar to the real contact data. Moreover, as the total number and duration of actual contacts cannot either be easily guessed from the co-presence data alone, we consider the actual total contact time $T_c$ as the (only) parameter of the sampling, and we fix it to its empirical value. The sampling methods we consider are the following:

- *Sampling 1: Sampling of co-presence times.* We define a co-presence list as a list of individuals present at the same time $t$ in the same area. Each co-presence list is thus stamped with its time of occurrence $t$. We create $n_\ell$ copies of each co-presence list $\ell$, where $n_\ell$ is the number of distinct individuals in $\ell$, and create in this way of a global pool of co-presence lists. We then sample $T_c$ lists uniformly at random from the pool without replacement. Each list has thus a probability proportional to the number of individuals it contains to be chosen. From each chosen list, we choose at random a pair $i, j$ of individuals, obtaining a triplet $(t, i, j)$ where $t$ is the time-stamp of the list (we take care to avoid repetitions: if $(t, i, j)$ has already be obtained in a previous random draw, we repeat the random selection). The sampled temporal co-presence network (i.e., the surrogate contact network) is formed by the union of these triplets. Note that this method does not conserve the durations of the co-presence events.

- *Sampling 2: Sampling of co-presence times with completion.* We constitute a pool of lists exactly like in the previous method. We then sample a triplet $(t, i, j)$ as in the previous method, and add all the other triplets $(t', i, j)$ that belong to the same co-presence event to create the surrogate contact event. We iterate this until we reach a cumulative contact time $T_c$, while discarding repetitions. Contrary to the previous sampling, this method conserves the durations of the co-presence events.
- *Sampling 3: Sampling of co-presence events.* We consider directly the list of co-presence events between individuals, $(t, i, j, \tau)$ (co-presence event between individuals $i$ and $j$, starting at time $t$ and with duration $\tau$), and sample events from this list, without replacement, adding them to the list of surrogate contact events until we reach a cumulative contact time $T_c$.

For each data set, we create 100 instances of surrogate contact networks for each sampling method. We compare in the following the properties of these surrogate contact networks with the real face-to-face contact data.

### 3.2 Network comparison

Figures 5–6 and Tables 2–3 provide elements of comparison between the surrogate contact networks and the empirical data (see also Additional file 1). The first observation is that the contact activity timelines are in general broadly recovered, except for the primary school (see Additional file 1), while the detailed intra-day activity variations are not always properly reconstructed in the surrogate data (except for the hospital data, see Additional file 1). The strongest deviations are observed for the second sampling method for the conference and high school data.

The first sampling method, given it samples separately times of co-presence, yields an exponential distribution of surrogate contact durations, in contrast with actual data and other sampling methods in which broad distributions are observed. Methods 2 and 3 generate broad distributions of the contact durations, either with an accurate slope or with a smaller exponent. For instance, the second sampling method systematically leads to a distribution of contact durations that is broader than for the real contacts. The third method yields a distribution of contact durations similar to the real one for the InVS13, LH10, and SFHH cases, but gives results similar to the second method in the other cases. Broad distributions of inter-contact durations and of the numbers of contacts between individuals can also be obtained for all methods, depending on the context, with slopes either accurate or smaller. Finally, distributions of link weights are usually rather well recovered by all methods.

We now turn to the properties of networks aggregated over daily periods or over the whole data collection. At the daily level, we show in Table 2 that the similarity of the contact matrices obtained from the surrogate data with the empirical one is very high, and most often larger than the similarity of the original co-presence matrix. For networks aggregated over the whole data collection, Fig. 5 shows the distributions of degrees and of weights (see also Additional file 1). The first sampling method leads to an overestimation of degree values (resulting in a shift of the distribution), the second method tends to shift the distribution to lower degree values (except for the conference case), and the third method yields context-dependent over- or under-estimations of degree values. Note that

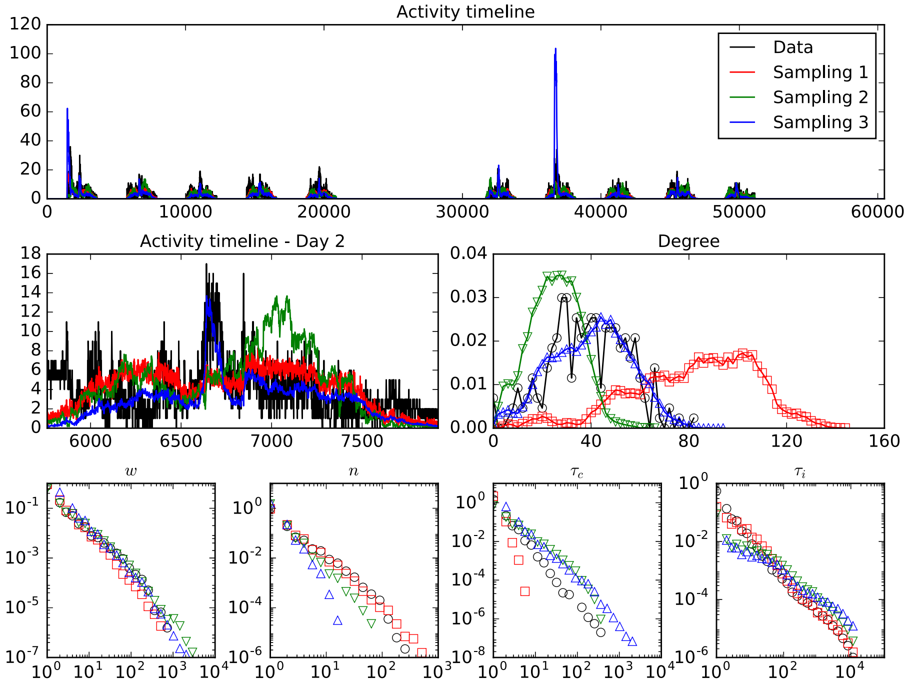

**Figure 5** Properties of the sampled co-presence networks—InVS15. We compare several properties of the contact network from the original data set with the surrogate contacts obtained by sampling of the co-presence data: overall timeline of contact activity, distributions of degree, weight $w$ and number of contacts per link $n$ in the network aggregated over the whole data collection period, and distributions of the contact duration $\tau_c$ and inter-contact duration $\tau_i$.

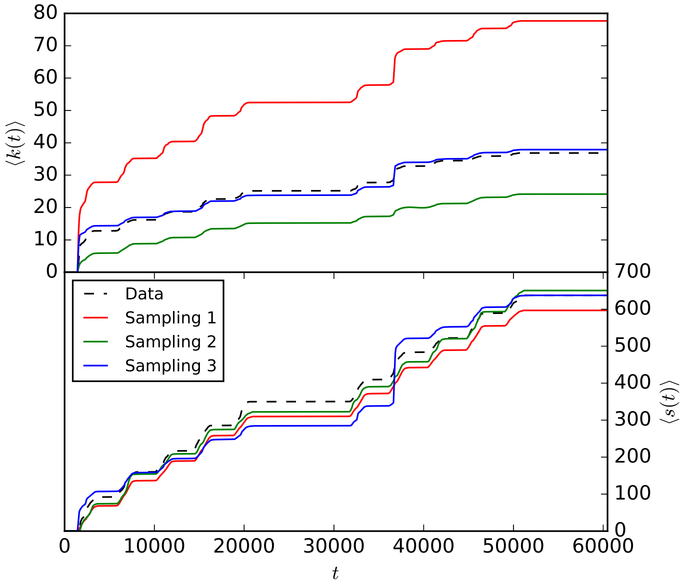

**Figure 6** Evolution of the mean aggregated degree and strength. InVS15 data set. We compare how the average degree $\langle k(t)\rangle$ and the average strength $\langle s(t)\rangle$ grow as we aggregate the network on increasing time-windows, for the real contact data and each sampling method.

**Table 4** Average similarity between daily networks

| | InVS13 | InVS15 | LH10 | LyonSchool | Thiers13 |
|---|---|---|---|---|---|
| Contact | 0.333 | 0.305 | 0.351 | 0.643 | 0.431 |
| Co-presence | 0.415 | 0.348 | 0.449 | 0.806 | 0.683 |
| Sampling 1 | 0.361 | 0.344 | 0.436 | 0.749 | 0.515 |
| Sampling 2 | 0.388 | 0.271 | 0.403 | 0.175 | 0.084 |
| Sampling 3 | 0.286 | 0.205 | 0.437 | 0.042 | 0.071 |
| Null model | 0.022 | 0.010 | 0.061 | 0.046 | 0.010 |

For each data set we compute the cosine similarity between the neighbourhoods of each nodes from each daily network, averaged for all nodes and all pairs of daily networks. The neighbourhood of a node $n$ is defined as the vector of the link weights between $n$ and every other nodes (if the link does not exist the weight is set to zero). We compare the values obtained for the contact data, the co-presence data, and for the networks generated by each sampling method of the co-presence data, averaged over $100$ realisations for each sampling method. For reference, we also compute as null model the average similarity when links in the contact data are shuffled randomly within each daily network.

the distributions of degrees of the co-presence networks are not shown in the figure as the degree values are very strongly overestimated. Distributions of weights (aggregated contact durations) recover well the ones of the data for all sampling methods, and are closer than the ones of the co-presence networks.

To investigate intermediate timescales of aggregation, Table 4 quantifies the similarity between networks aggregated in different days. The measure is defined as the average cosine similarity between all pairs of instances of a node's neighbourhood, averaged over all nodes. We see that the similarity is higher for the co-presence networks, as expected since the networks are denser. The sampling method 1 generates networks that are more similar than the data, and the other two methods generate networks that are less similar (with the exception of the LH10 case, and the method 2 in the InVS13 case). In the cases of the method 2 for the LyonSchool data, and the methods 2 and 3 for the Thiers13 data, the sampled networks are even almost as different as they would be after a random shuffling of the links.

In addition, Fig. 6 gives the evolution of the average degree and strength for networks aggregated in increasingly long time windows. First, the evolution of the real average aggregated strength is usually better recovered than for the degree by the various sampling better. Second, which sampling method recovers better the evolution of the degree is again context dependent. However, in all cases the sampled data are much closer to the contact data than the co-presence network, which overestimates very strongly these quantities.

### 3.3 Node centralities

In a network, more "central" nodes are usually considered as important, as they might play an important role for instance in spreading processes (or other dynamical phenomena) occurring in the network. It is thus of interest to understand whether the most central nodes in the contact network can be identified either in the raw co-presence data or in the surrogate contact data built from the co-presence information. As there are several ways of determining central nodes in a network, we consider here three of the most well-known centrality measures and apply them to the networks aggregated over the whole data collection: degree $k$, strength $s$ and betweenness $b$ of nodes in the aggregated networks. For each instance of each sampling method, we thus build the resulting surrogate aggregated contact network and rank nodes according to each centrality measure. We then compute the Jaccard similarity index between the top $N\%$ nodes in the real contact network and in

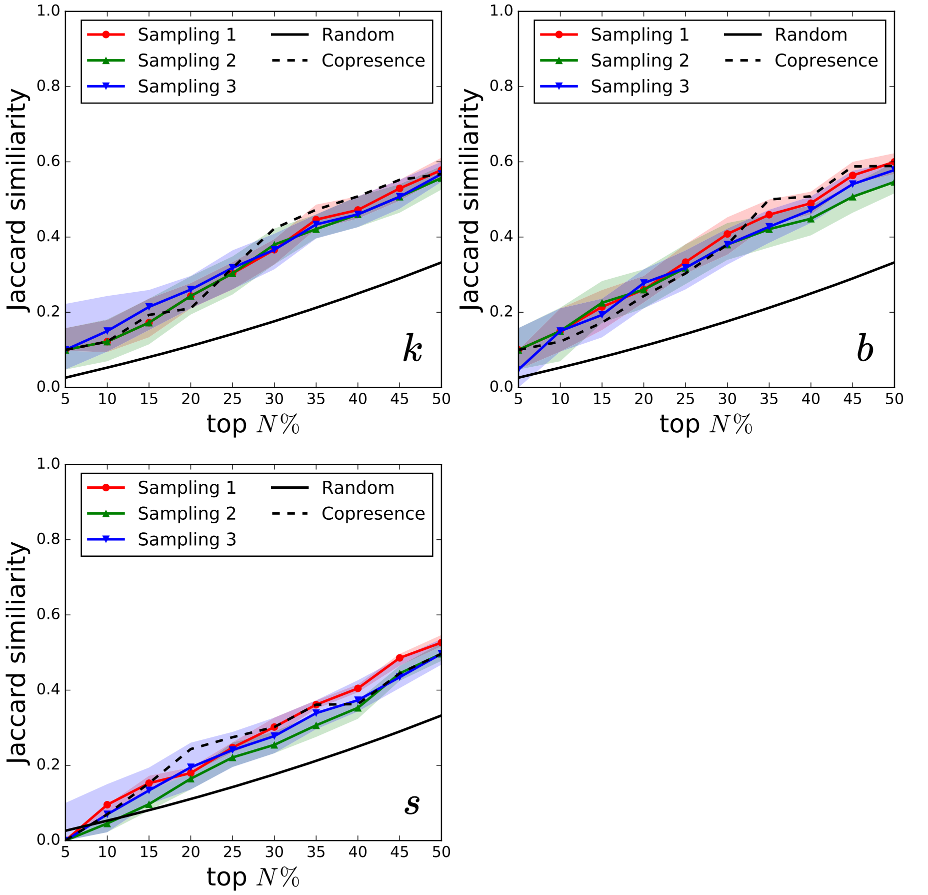

**Figure 7** Node ranking similarity. InVS15 data set. We plot for each co-presence sampling method the Jaccard similarity between the top $N$% nodes in the real and surrogate contact data, when ranked according to their degree $k$, their strength $s$ or their betweenness centrality $b$ vs. $N$. The plot shows the median similarity and the shaded areas give the 90% confidence interval.

the surrogate one. We plot in Fig. 7 the median similarity with the 90% confidence interval, as a function of $N$, for the InVS15 case (see Additional file 1 for the other cases).

In general, no sampling method recovers correctly the most central nodes for low values of $N$. The best results are obtained for the conference data with similarities around 0.2–0.4. The similarity values increase as $N$ increases but reach most often only values of $\sim 0.5$ when considering the top 50% nodes, meaning that only 25% of the most central nodes are identified when using the surrogate data. The best results are obtained for the first sampling method for the LyonSchool case and for the LH10 case, with similarities reaching 0.6–0.7. Results are typically better than the random baseline but do not outperform the detection of most central nodes based on the whole co-presence network. In terms of the most central nodes as defined by the $k$-core decomposition (we recall that the $k$-core of a network is the maximal subgraph such that all nodes in the subgraph have at least degree $k$, and $k$ is called the coreness), the overestimation of degrees in the co-presence network leads to an overestimation of the maximum coreness, while sampling leads to values closer to the ones of the contact data, but once again in a context-dependent way. The maximum core itself is only partially recovered in the whole and in the sampled co-presence networks (see Table 5).

**Table 5** Comparison of the maximum *k*-core properties

| | InVS13 | InVS15 | LH10 | LyonSchool | SFHH | Thiers13 |
|---|---|---|---|---|---|---|
| Contact | 11 | 25 | 23 | 47 | 33 | 24 |
| Co-presence | 78 (0.607) | 112 (0.681) | 32 (0.682) | 181 (0.615) | 320 (0.522) | 210 (0.684) |
| Sampling 1 | 22.5 (0.660) | 57.8 (0.719) | 26.0 (0.683) | 99.8 (0.638) | 111 (0.575) | 76.7 (0.640) |
| Sampling 2 | 5.23 (0.479) | 17.4 (0.692) | 16.4 (0.655) | 39.7 (0.501) | 41.4 (0.555) | 17.7 (0.375) |
| Sampling 3 | 28.1 (0.591) | 27.4 (0.639) | 25.5 (0.693) | 28.4 (0.360) | 42.6 (0.559) | 15.8 (0.117) |

For each dataset we compute the maximum coreness, and report between parenthesis the Jaccard index between the k-core of the contact network and the k-core in the original and sampled co-presence data (results are averaged over 100 realisations for each sampling method).

## 4 Using surrogate contact data in epidemic simulations

We have seen in the previous section that none of the three sampling methods yields a perfectly accurate description of all the relevant features of the true contact network: each sampling method yields surrogate data with both interesting similarities and potentially important discrepancies with respect to the original contact data. We now consider the issue of using such surrogate data in simulations of spreading processes: as precise data on face-to-face contacts is not always available, it is important to understand if co-presence information can allow us to obtain on the one hand an accurate prediction of the outcome of an epidemic process, and on the other hand a reliable estimation of the impact of containment measures. In particular, it is important to be able to classify potential containment strategies to determine which one(s) are most adequate.

To this aim, we consider the paradigmatic Susceptible-Infectious-Recovered (SIR) model for epidemic spreading. In this model, susceptible (S) individuals can become infectious (I) at rate $\beta$ when in contact with an infectious node. Infectious nodes recover spontaneously at rate $\mu$ and enter an immune *recovered* (R) state. Simulations start with a single infectious individual chosen at random and carried out until there are no infectious individuals left in the population, i.e., individuals are either still susceptible or have been infectious and have then recovered. The impact of the epidemics is then quantified by the final fraction $n_i$ of individuals in the R state.

We set $\beta = 0.0004$ (corresponding to an average infection time of 2,500 s) and vary $\mu$ by tuning the reproductive number $R = \beta/\mu$. For each value of $R$, we measure the fraction $P(n_i > 20\%)$ of "large" outbreaks in which the fraction $n_i$ of the population that was reached by the outbreak is at least 20% and the distribution of the sizes $n_i$ of these large outbreaks. We average the results over 10,000 simulations performed on the empirical contact network. For each sampling method, we build 100 different instances of the surrogate contact network, and perform 100 simulations on each surrogate network.

We also consider several simple methods to mitigate the spread, namely the vaccination of a number of individuals in the population, under the assumption of a perfect vaccine efficiency: vaccinated individuals cannot become infectious nor transmit the disease and thus slow down and hinder the propagation. We consider the vaccination of (i) 5, 10 or 20 individuals chosen at random (ii) the most central 5, 10 or 20 individuals, where centrality is measured according to either degree, strength or betweenness in either the real or surrogate contact networks (iii) when the population is structured in groups, the vaccination of all individuals in one group.

Figures 8 and 9 summarize our results for the InVS15 dataset (see Additional file 1 for the figures obtained with the other datasets). In terms of the evaluation of the impact of

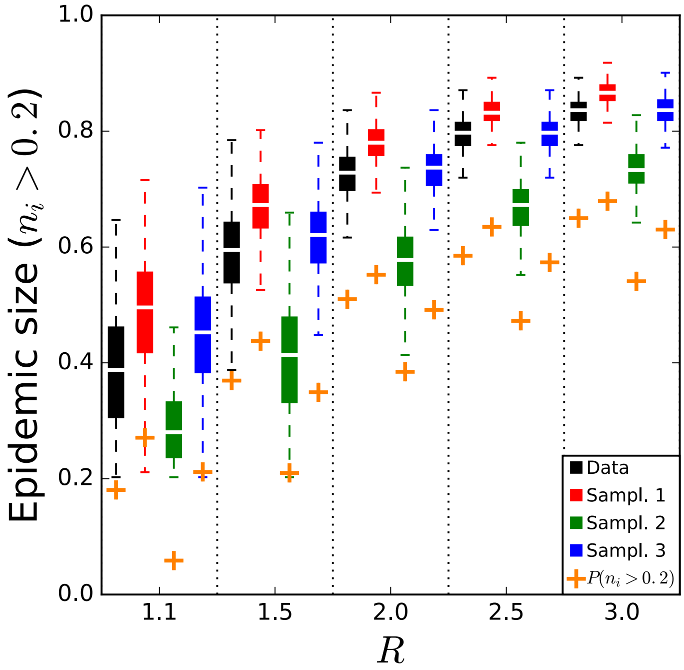

**Figure 8** Epidemic prevalence. InVS15 data set. We plot the fraction of the total number of outbreaks that reach at least 20% of the population (crosses) and the distribution of the sizes of these outbreaks (boxplots) for several values of the reproductive number *R*, for the original and the surrogate contact networks.

a spreading process, results are context dependent. The simulations performed on the surrogate data obtained with the first method generally lead to an overestimation of the epidemic risk, except for the hospital data. When using the second sampling method, we obtain a good estimation of the risk for the conference, school and highschool data but an underestimation for offices and hospital data. The third method on the other hand leads to a correct estimation for the offices and hospital data but an underestimation for the school and highschool and an overestimation for the conference.

We show in Fig. 9 the impact of the various vaccination strategies, quantified through the ratio of the probabilities of large outbreaks with and without vaccination, as well as the ratio between the median sizes of these large outbreaks. We rank the strategies according to their efficiency in the real contact network, in order to visualize easily whether the surrogate networks lead to the same classification of the strategies: indeed, even when the impact of each specific strategy is not accurately quantified, it would be interesting at least to understand which methods are most efficient.

Results are once again uneven and context dependent (see also Table 6). In several cases such as SFHH the ranking of strategies obtained from the sampled co-presence is overall respected (Kendall's tau of 0.818 for the sampling method 1 on the size of outbreaks), while it can be strongly reshuffled in other cases (for instance in the Thiers13 case).

## 5 Discussion and conclusion

In this paper, we have investigated whether low resolution co-presence information can be used as a substitute for detailed face-to-face proximity data, both from the point of view of extracting large-scale structural and statistical features of the temporal contact

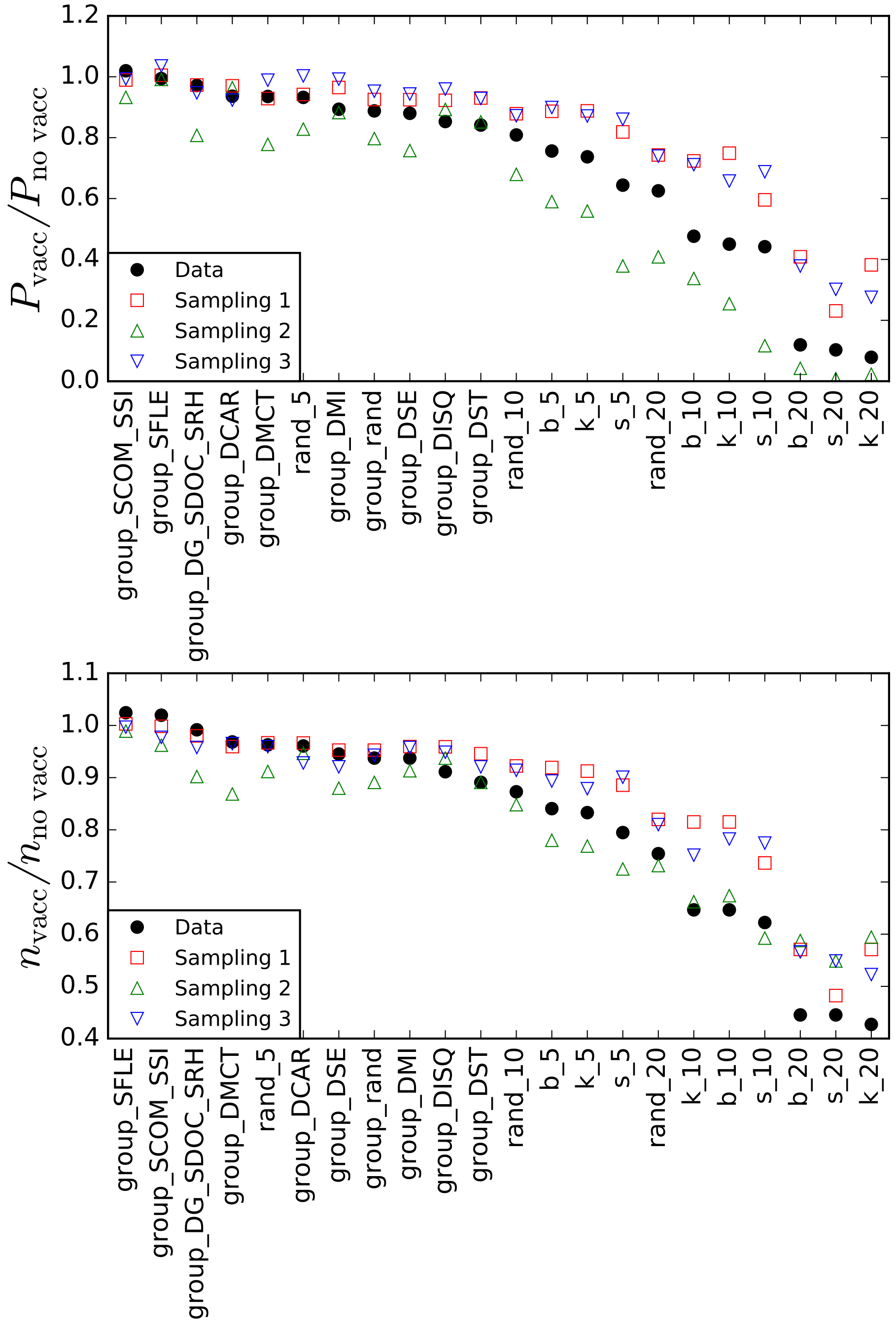

**Figure 9** Vaccination strategies. InVS15 data set. We plot the ratio between the vaccination and no vaccination cases of the fraction of the total number of outbreaks that reach at least 20% of the population (top), and of the median size of these outbreaks (bottom) for different vaccination strategies, for the original data and the reconstructed networks. The vaccination strategies are ordered by decreasing efficiency, based on the effect on the real contact data. The *group_** strategies consist in vaccinating one or several groups entirely; the *group_rand* strategy vaccinates $n_g$ random nodes, where $n_g$ is the average group size; the *rand_n* strategies randomly vaccinates a specified fraction *n* of nodes; the *b_n*, *k_n*, *s_n* strategies vaccinate the top *n*% nodes according to, respectively, betweenness centrality, degree and strength ranking.

**Table 6** Comparison of the vaccination strategy rankings

| | InVS13 | | InVS15 | | LH10 | | LyonSchool | | SFHH | | Thiers13 | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | *P* | *n* | *P* | *n* | *P* | *n* | *P* | *n* | *P* | *n* | *P* | *n* |
| Sampling 1 | 0.515 | 0.235 | 0.377 | 0.766 | 0.397 | 0.412 | –0.012 | –0.051 | 0.485 | 0.818 | 0.048 | 0.299 |
| Sampling 2 | 0.324 | 0.382 | 0.377 | 0.481 | 0.456 | 0.412 | 0.091 | 0.083 | 0.758 | 0.485 | –0.074 | –0.108 |
| Sampling 3 | 0.279 | 0.235 | 0.394 | 0.706 | 0.324 | 0.426 | 0.020 | 0.162 | 0.636 | 0.545 | 0.299 | –0.108 |

For each sampling method we compute Kendall's tau between the list of vaccination strategies ranked by increasing efficiency for the contact data and for the sampled co-presence networks, both in terms of the fraction of large outbreaks ($P$) and of median sizes of the large outbreaks ($n$).

network in a population and in data-driven models of epidemic processes in a population. We have considered several data sets collected in various contexts that contain both high-resolution data on face-to-face contacts between individuals and a coarser location data, both with temporal resolution. The location data can thus be transformed into a co-presence temporal network between individuals. Given its lower spatial resolution, this co-presence data contains much more events than the contact data, leading to much denser aggregated networks: indeed, all individuals in a given area are considered as co-present, while only some of them are typically engaged in a face-to-face contact. Despite this expected issue, a number of properties related to group structure and statistical distributions of temporal properties are similar in contact and co-presence data, with similar matrices of densities of links between groups and broad distributions of (aggregate) contact durations.

We have thus examined several methods to downsample the co-presence networks to create surrogate contact networks with overall the same amount of contact time than the real contact data. The surrogate data statistics are in general closer to the real contact data than the raw co-presence, in particular regarding the distribution of node degrees and link weights (and their evolution in networks aggregated over increasing time windows). These results mean in particular that the distribution of aggregate contact durations, a very important property that has a strong impact on the unfolding of processes on networks such as epidemic processes, could be approximately retrieved from simple sampling processes of the co-presence data and thus fed into data-driven models of populations. Several other properties, such as precise value of the average degree, average clustering or size of largest cliques and cores, turn out however to be strongly context-dependent. Moreover, the most central nodes of the contact network are not better identified than using the bare co-presence information.

We have moreover investigated the use of such surrogate contact data in numerical simulations of spreading processes in a population. Overall, simulations performed on surrogate data obtained with one of the sampling method yield results close to the ones obtained with the real data, while the other methods over- or under-estimate these results, but the best method turns out to depend on context (Note however that all these methods give obviously results much closer to the one of the real contact network than if raw co-presence is used, given co-presence overestimates strongly the contacts and thus yields a strongly overestimated epidemic risk). We moreover investigated the possibility to rank containment strategies according to their efficiency, and found that this ranking is once again context dependent: in some cases, simulations on sampled co-presence networks allow us to uncover the most efficient vaccination strategies for containing a spread on the real contact data, while in other cases the rankings differ quite strongly.

In conclusion, we showed that co-presence data, while yielding interesting insights into some of the large scale properties of the contact network, is not easily usable to build in a reliable and systematic fashion surrogate contact data that reproduces detailed features of the real contacts and could be used in numerical simulations to predict the outcome of spreading processes and the impact of containment strategies, at least for processes involving contagion at short distances [39] (note that, while more sophisticated sampling procedures might be devised, they would most probably involve more parameters and/or more additional information not present in the raw co-presence data, and would also most probably still give context-dependent results). The SocioPatterns data that was used is representative of the current state of the art in data collection of human behaviour, both for face-to-face contact and localisation. The results and methods presented in the article could thus be easily applied to data generated by any other type of system, including Bluetooth proximity sensing or WiFi tracking. We finally note that even coarse location information has been shown to be a useful additional information whenever the precise contact data is incomplete [29]. Optimally, data collection with wearable sensors should thus contain both high resolution data about relative positions of individuals, in order to detect face-to-face proximity, and coarser co-presence information to inform for instance on mobility patterns within buildings or complement potential data losses.

## Additional material

**Additional file 1:** This file contains additional tables and figures, in particular for the other data sets considered in this paper. (PDF 3.5 MB)

**Funding**
The publication of this article was funded by the Open Access Fund of the Leibniz Association.

**Availability of data and materials**
The data used for the present paper is available at: https://zenodo.org/record/1117884. The files are the following: *contact.tar.bz2* contains all six contact networks, *co-presence.tar.bz2* contains all six co-presence networks. Both contacts and co-presence data are formatted as *tij*, i.e. each line represents a contact occurring at a time *t* between two nodes *i* and *j*. *metadata.tar.bz2* contains the lists of nodes, with the first column being the node identifier and the second the group affiliation, when available.

**Competing interests**
The authors declare that they have no competing interests.

**Authors' contributions**
MG and AB conceived and designed the study. MG performed the numerical simulations and statistical analysis, created the figures and wrote the first draft of the manuscript. MG and AB wrote, read and approved the final version of the manuscript.

**Author details**
[1]CNRS, CPT, Aix Marseille Univ, Université de Toulon, Marseille, France. [2]GESIS, Leibniz Institute for the Social Sciences, Köln, Germany. [3]Data Science Laboratory, ISI Foundation, Torino, Italy.

## Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Received: 21 December 2017 Accepted: 25 April 2018 Published online: 08 May 2018

**References**

1. Mossong J, Hens N, Jit M, Beutels P, Auranen K, Mikolajczyk R, Massari M, Salmaso S, Tomba GS, Wallinga J, Heijne J, Sadkowska-Todys M, Rosinska M, Edmunds WJ (2008) Social contacts and mixing patterns relevant to the spread of infectious diseases. PLoS Med 5(3):e74. https://doi.org/10.1371/journal.pmed.0050074
2. Danon L, Read J, House T, Vernon M, Keeling M (2013) Social encounter networks: characterizing Great Britain. Proc - Royal Soc, Biol Sci 280:20131037. https://doi.org/10.1098/rspb.2013.1037

3. Eames K, Bansal S, Frost S, Riley S (2015) Six challenges in measuring contact networks for use in modelling. Epidemics 10:72–77
4. Eagle N, (Sandy) Pentland A (2006) Reality mining: sensing complex social systems. Pers Ubiquitous Comput 10(4):255–268. https://doi.org/10.1007/s00779-005-0046-3
5. O'Neill E, Kostakos V, Kindberg T, Schiek A, Penn A, Fraser D, Jones T (2006) Instrumenting the city: developing methods for observing and understanding the digital cityscape. In: Dourish P, Friday A (eds) UbiComp 2006. Ubiquitous Computing. Lecture notes in computer science, vol 4206. Springer, Berlin, pp 315–332. https://doi.org/10.1007/11853565_19
6. Scherrer A, Borgnat P, Fleury E, Guillaume J-L, Robardet C (2008) Description and simulation of dynamic mobility networks. Comput Netw 52(15):2842–2858
7. Vu L, Nahrstedt K, Retika S, Gupta I (2010) Joint bluetooth/WiFi scanning framework for characterizing and leveraging people movement in university campus. In: Proceedings of the 13th ACM international conference on modeling, analysis, and simulation of wireless and mobile systems (MSWIM'10). ACM, New York, NY, USA, pp 257–265. https://doi.org/10.1145/1868521.1868563
8. Zhang Y, Wang L, Zhang Y-Q, Li X (2012) Towards a temporal network analysis of interactive WiFi users. Europhys Lett 98(6):68002
9. Stopczynski A, Sekara V, Sapiezynski P, Cuttone A, Madsen MM, Larsen JE, Lehmann S (2014) Measuring large-scale social networks with high resolution. PLoS ONE 9(4):1–24. https://doi.org/10.1371/journal.pone.0095978
10. Olguín DO, Pentland AS (2008) Social sensors for automatic data collection. In: Proceedings of the fourteenth Americas conference on information systems (AMCIS 2008), pp 171
11. Hashemian MS, Stanley KG, Osgood N (2010) Flunet: automated tracking of contacts during flu season. In: 8th international symposium on modeling and optimization in mobile, ad hoc, and wireless networks, pp 348–353
12. Salathé M, Kazandjieva M, Lee JW, Levis P, Feldman MW, Jones JH (2010) A high-resolution human contact network for infectious disease transmission. Proc Natl Acad Sci 107(51):22020–22025. http://www.pnas.org/content/107/51/22020.full.pdf. https://doi.org/10.1073/pnas.1009094108
13. Cattuto C, Van den Broeck W, Barrat A, Colizza V, Pinton J, Vespignani A (2010) Dynamics of person-to-person interactions from distributed RFID sensor networks. PLoS ONE 5(7):11596. https://doi.org/10.1371/journal.pone.0011596
14. Berke EM, Choudhury T, Ali S, Rabbi M (2011) Objective measurement of sociability and activity: mobile sensing in the community. Ann Family Med 9(4):344–350. http://www.annfammed.org/content/9/4/344.full.pdf+html. https://doi.org/10.1370/afm.1266
15. Lucet J-C, Laouenan C, Chelius G, Veziris N, Lepelletier D, Friggeri A, Abiteboul D, Bouvet E, Mentre F, Fleury E (2012) Electronic sensors for assessing interactions between healthcare workers and patients under airborne precautions. PLoS ONE 7(5):1–7. https://doi.org/10.1371/journal.pone.0037893
16. Hornbeck T, Naylor D, Segre AM, Thomas G, Herman T, Polgreen PM (2012) Using sensor networks to study the effect of peripatetic healthcare workers on the spread of hospital-associated infections. J Infect Dis 206(10):1549–1557. Accessed 2017-09-28. https://doi.org/10.1093/infdis/jis542
17. Lowery-North DW, Hertzberg VS, Elon L, Cotsonis G, Hilton SA, Vaughns CF II, Hill E, Shrestha A, Jo A, Adams N (2013) Measuring social contacts in the emergency department. PLoS ONE 8(8):1–9. https://doi.org/10.1371/journal.pone.0070854
18. Toth DJA, Leecaster M, Pettey WBP, Gundlapalli AV, Gao H, Rainey JJ, Uzicanin A, Samore MH (2015) The role of heterogeneity in contact timing and duration in network models of influenza spread in schools. J R Soc Interface 12(108):20150279. https://doi.org/10.1098/rsif.2015.0279
19. Guclu H, Read J, Vukotich CJ Jr, Galloway DD, Gao H, Rainey JJ, Uzicanin A, Zimmer SM, Cummings DAT (2016) Social contact networks and mixing among students in K-12 schools in Pittsburgh, PA. PLoS ONE 11(3):1–19. https://doi.org/10.1371/journal.pone.0151139
20. Blower S, Go M-H (2011) The importance of including dynamic social networks when modeling epidemics of airborne infections: does increasing complexity increase accuracy? BMC Med 9(1):88. https://doi.org/10.1186/1741-7015-9-88
21. Barrat A, Cattuto C, Tozzi AE, Vanhems P, Voirin N (2014) Measuring contact patterns with wearable sensors: methods, data characteristics and applications to data-driven simulations of infectious diseases. Clin Microbiol Infect 20(1):10–16. https://doi.org/10.1111/1469-0691.12472
22. De Choudhury M, Mason WA, Hofman JM, Watts DJ (2010) Inferring relevant social networks from interpersonal communication. In: Proceedings of the 19th international conference on World Wide Web (WWW'10). ACM, New York, NY, USA, pp 301–310. https://doi.org/10.1145/1772690.1772722
23. Eagle N, Pentland AS, Lazer D (2009) Inferring friendship network structure by using mobile phone data. Proc Natl Acad Sci 106(36):15274–15278. http://www.pnas.org/content/106/36/15274.full.pdf. https://doi.org/10.1073/pnas.0900282106
24. Crandall DJ, Backstrom L, Cosley D, Suri S, Huttenlocher D, Kleinberg J (2010) Inferring social ties from geographic coincidences. Proc Natl Acad Sci 107(52):22436–22441. http://www.pnas.org/content/107/52/22436.full.pdf. https://doi.org/10.1073/pnas.1006155107
25. Scholz C, Atzmueller M, Stumme G, Barrat A, Cattuto C (2013) New insights and methods for predicting face-to-face contacts. In: 7th international conference on weblogs and social media, ICWSM-13
26. Sapiezynski P, Stopczynski A, Wind DK, Leskovec J, Lehmann S (2017) Inferring person-to-person proximity using WiFi signals. Proc ACM Interact Mob Wearable Ubiquitous Technol 1(2):24–12420. https://doi.org/10.1145/3090089
27. SocioPatterns Collaboration. www.sociopatterns.org. Accessed 18 Oct 2017
28. Génois M, Vestergaard CL, Cattuto C, Barrat A (2015) Compensating for population sampling in simulations of epidemic spread on temporal contact networks. Nat Commun 6:8860. https://doi.org/10.1038/ncomms9860
29. Sapienza A, Barrat A, Cattuto C, Gauvin L (2017) Estimating the outcome of spreading processes on networks with incomplete information: a mesoscale approach. https://arxiv.org/abs/1709.01806
30. Smieszek T, Salathé M (2013) A low-cost method to assess the epidemiological importance of individuals in controlling infectious disease outbreaks. BMC Med 11(1):35. https://doi.org/10.1186/1741-7015-11-35

31. Génois M, Vestergaard CL, Fournet J, Panisson A, Bonmarin I, Barrat A (2015) Data on face-to-face contacts in an office building suggest a low-cost vaccination strategy based on community linkers. Netw Sci 3:326–347. https://doi.org/10.1017/nws.2015.10
32. Vanhems P, Barrat A, Cattuto C, Pinton J-F, Khanafer N, Régis C (2013) Estimating potential infection transmission routes in hospital wards using wearable proximity sensors. PLoS ONE 8(9):73970. https://doi.org/10.1371/journal.pone.0073970
33. Stehlé J, Voirin N, Barrat A, Cattuto C, Isella L, Pinton J-F, Quaggiotto M, Van den Broeck W, Régis C, Lina B, Vanhems P (2011) High-resolution measurements of face-to-face contact patterns in a primary school. PLoS ONE 6(8):23176. https://doi.org/10.1371/journal.pone.0023176
34. Isella L, Stehlé J, Barrat A, Cattuto C, Pinton J-F, Van den Broeck W (2011) What's in a crowd? Analysis of face-to-face behavioral networks. J Theor Biol 271(1):166–180. https://doi.org/10.1016/j.jtbi.2010.11.033
35. Mastrandrea R, Fournet J, Barrat A (2015) Contact patterns in a high school: a comparison between data collected using wearable sensors, contact diaries and friendship surveys. PLoS ONE 10(9):1–26. https://doi.org/10.1371/journal.pone.0136497
36. Stehlé J, Voirin N, Barrat A, Cattuto C, Colizza V, Isella L, Régis C, Pinton J-F, Khanafer N, Van den Broeck W, Vanhems P (2011) Simulation of an seir infectious disease model on the dynamic contact network of conference attendees. BMC Med 9(1):87. https://doi.org/10.1186/1741-7015-9-87
37. Schläpfer M, Bettencourt LM, Grauwin S, Raschke M, Claxton R, Smoreda Z, West GB, Ratti C (2014) The scaling of human interactions with city size. J R Soc Interface 11:20130789
38. Tizzoni M, Sun K, Benusiglio D, Karsai M, Perra N (2015) The scaling of human contacts and epidemic processes in metapopulation networks. Sci Rep 5:15111
39. Stopczynski A, Pentland A, Lehmann S (2015) Physical proximity and spreading in dynamic social networks. https://arxiv.org/abs/1509.06530