Sparking "The BBC Four Pandemic": Leveraging citizen science and mobile phones to model the spread of disease

Stephen M. Kissler[a*], Petra Klepac[b,a,c], Maria Tang[a], Andrew J.K. Conlan[d], Julia R. Gog[a]

[a] *Department of Applied Mathematics and Theoretical Physics, University of Cambridge, Wilberforce Road, Cambridge (UK), CB3 0WA*

[b] *Department of Infectious Disease Epidemiology, Faculty of Epidemiology and Public Health, London School of Hygiene and Tropical Medicine, Keppel Street, London (UK), WC1E 7HT*

[c] *Centre for Mathematical Modelling of Infectious Diseases, London School of Hygiene and Tropical Medicine, Keppel Street, London (UK), WC1E 7HT*

[d] *Department of Veterinary Medicine, University of Cambridge, Madingley Road, Cambridge (UK), CB3 0ES*

[*] *To whom correspondence should be addressed. Email: sk792@cam.ac.uk*

November 6, 2018

Keywords: Human mobility | Disease transmission | Realistic interaction patterns | Dynamic network | Vaccination optimisation

Abstract

The nexus of mobile technology, mass media, and public engagement is opening new opportunities for research into the human behaviours relevant to the spread of disease. On 22 March 2018, the British Broadcasting Corporation (BBC) released the documentary "Contagion! The BBC Four Pandemic" to describe the science behind pandemic preparedness in the UK. The authors of this article were responsible for producing a mathematical simulation of how a highly contagious respiratory pathogen might spread across the UK. According to the documentary narrative, the 'outbreak' begins in the town of Haslemere, England. To ground the simulation in true human interaction patterns, a three-day citizen science experiment was conducted during which the pairwise distances between 469 volunteers in Haslemere were tracked continuously using a mobile phone app. Here, we offer a scientific companion to the documentary in which we describe the methods behind our simulation and release the pairwise distance dataset. We discuss salient features of the dataset, including daily patterns in the clustering and volatility of interpersonal interactions. We also discuss new scientific opportunities opened by the 'Haslemere dataset', such as the ability to compare rival methods for empirically estimating the basic reproduction number and the opportunity to develop optimal vaccination strategies. We believe that the Haslemere dataset will productively challenge current strategies for incorporating population structure into disease transmission models, and hope that it will inspire the collection and analysis of other similar datasets in the future.

1 Introduction

On 22 March 2018, the television channel BBC Four released "Contagion! The BBC Four Pandemic" [1], a documentary on influenza epidemiology to mark the centenary of the 1918 Spanish Flu pandemic. The documentary explores how a highly contagious new strain of influenza could spread across the United Kingdom using a simulated outbreak based on mathematical models developed by the authors of this article. According to the documentary narrative, a simulated outbreak originates in Haslemere, a town in Surrey, in the south of England, and then goes on to infect the rest of the UK. Here, we document the model used to produce the "Haslemere outbreak" described in the documentary, and we present the underlying data (the "Haslemere dataset") collected from volunteers in the town. A description of the UK-wide model is given by Klepac *et al.* (2018) [2].

The Haslemere outbreak simulation was produced in real time during a three-day data collection period. Participants downloaded the *BBC Pandemic* mobile phone app and then went about their daily business, with the app running in the background. The app collected records of the participants' GPS locations over time, which were sent nightly to the authors of this article to develop and test a disease transmission model. Following the final data pull, we simulated a single outbreak from our model to feature in the BBC Four documentary. That afternoon, a community event was held in Haslemere at which participants received a notification on their phones telling them whether or not they had been infected, and if they had, how many others they had gone on to infect. The Haslemere dataset, released with this article, is a fully anonymised version of the data collected during those three days.

Previous studies have used mobile phones and other tracking devices to uncover human mobility patterns relevant to the spread of disease; these may be roughly separated into studies of bulk movement [3, 4, 5, 6, 7, 8, 9, 10] and studies of interpersonal interactions [11, 12, 13, 14]. The Haslemere dataset falls into the latter category, and to our knowledge it is the most comprehensive such dataset yet collected, in terms of its geographic scope and number of volunteers.

This dataset also challenges what general types of models might best capture the human interactions relevant to the spread of disease. Networks have become a popular way of incorporating population structure into models of infectious disease transmission [15, 16, 17, 18, 19, 20, 21, 22].

Static, unweighted networks are particularly amenable to mathematical analysis, and have featured in a range of disease transmission models [21, 23, 22, 24, 25, 26, 27]. However, there is a growing body of evidence that suggests that this type of network may not adequately capture information about the frequency, proximity, and timing of interpersonal encounters, critical to the spread of disease. Interest in dynamic networks is increasing, both from theoretical [28, 29, 30] and empirical [31, 32, 33, 34] standpoints. The Haslemere dataset may help ground future dynamic network models in reality, and may reveal what information is lost when the the rich dynamics of human interactions are approximated by a static network.

The Haslemere dataset opens a wealth of opportunities for examining disease transmission and human behaviour in general. While we focus primarily on the work carried out for the BBC Four documentary, we also touch briefly upon other aspects of disease transmission that might be examined using the Haslemere dataset and others like it. Following Lloyd-Smith *et al.* (2005) [35], we characterise the superspreaders of the Haslemere "outbreak", and test whether the superspreaders of similar outbreaks might be predicted from the pairwise distance data alone. We also address the difficulty of estimating the basic reproduction number R_0 empirically, and highlight disagreements between different methods. We conclude with a discussion of optimal vaccination strategies, an area of research that may especially profit from datasets like the one presented here.

2 Description of the Haslemere dataset

The Haslemere dataset (Supplemental Data 1) consists of the pairwise distances of up to 1m resolution between 469 volunteers from Haslemere, England, at five-minute intervals over three consecutive days (Thursday 12 Oct – Saturday 14 Oct, 2017), excluding the hours between 11pm and 7am. Full details on the study site, data collection, and post-processing are provided in the Supplemental Materials and Methods. Fig. 1 depicts the pairwise encounters within 20m for a subset of the volunteers as a network, where each edge is coloured according to the quarter-day in which the encounter occurred [36]. This illustrates the heterogeneity and temporal complexity of the network, with some clustered individuals who encounter each other frequently, and other individuals who provide infrequent links between distant parts of the network. Relatedly, Movie

S1 depicts the pairwise interpersonal probabilities of infection for a subset of volunteers under the Haslemere outbreak transmission model (see Supplemental Materials and Methods), at four different temporal resolutions. The video illustrates how some connections persist over time, while other dynamic elements may be lost when the network is temporally integrated.

Mean degree, clustering, and link volatility

To illustrate the rich temporal structure present in the Haslemere dataset, we calculate the mean degree, mean local clustering coefficient [37], and mean link volatility [38] over the three-day study period (Fig. 2) using network 'snapshots' generated from the pairwise distance data. For each time point in the dataset, corresponding to a five-minute interval in real time, a network is generated in which each node represents an individual, and a link is defined between two nodes if the corresponding individuals are within some cutoff distance of one another during that five-minute interval. The mean link volatility $1 - \gamma$, introduced by Clauset and Eagle (2012) [38], extends the Pearson correlation coefficient to measure the persistence (γ) of network links between subsequent snapshots (see Materials and Methods). The mean degree and mean clustering coefficient are highest during the nighttime hours, while the mean link volatility is highest during the day (Fig. 2). The effect is most pronounced when the link-defining cutoff distance is 50m (red), but is still apparent for cutoff distances of 20m (purple) and 5m (blue). This may be interpreted as evidence that people tend to form temporally persistent small groups in the evenings, perhaps as families occupying the same household, and tend to have more fleeting interactions during the day with a wider range of people. This contrasts with Clauset and Eagle [38], who find in a study of the interpersonal interactions of students and professors at the Massachusetts Institute of Technology that the mean degree, mean local clustering coefficient, and mean volatility are all highest during the day. This difference may be attributed to the fact that our study captures the movements of people across a whole town, and thus represents a more general cross-section of the population, capturing interactions both outside and inside the home.

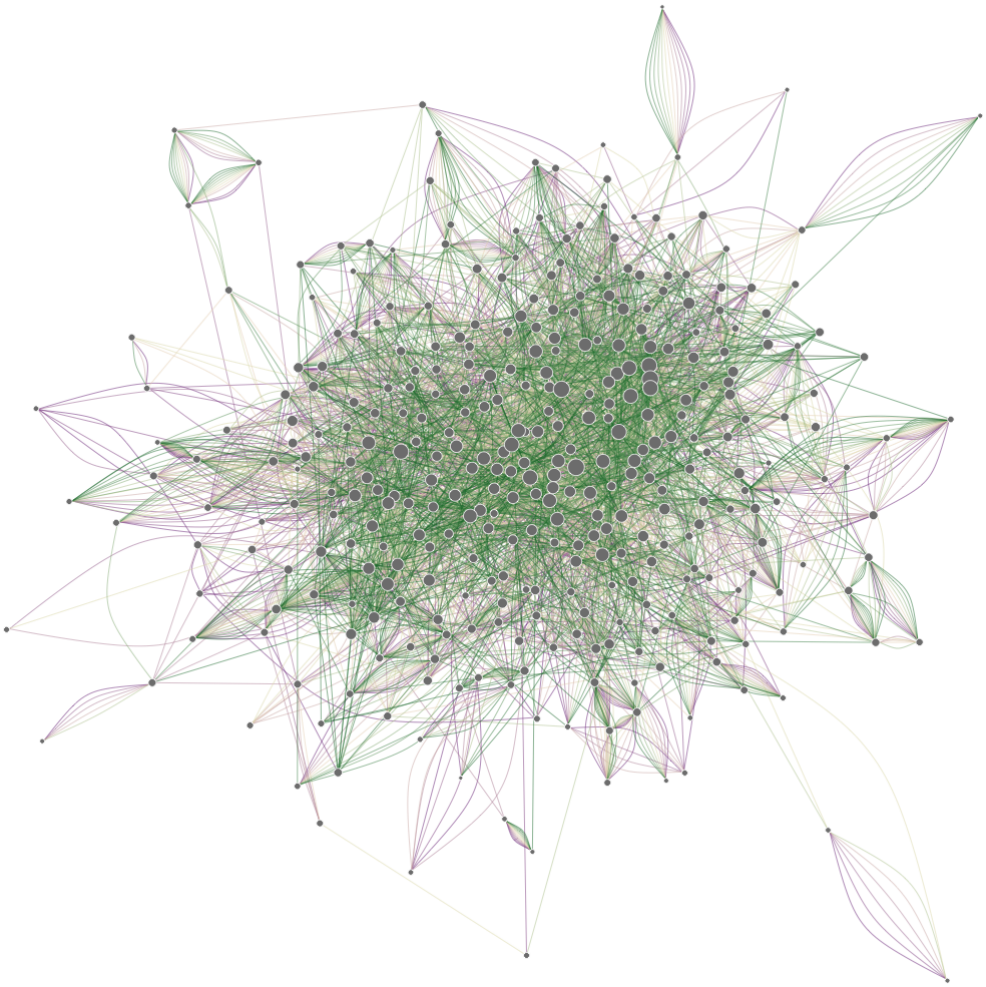

Figure 1: Network depicting pairwise encounters over time for the 75% of individuals in the Haslemere dataset who spend the greatest amount of time within 20m of another person (i.e. with the most total 'person-hours' of contact). An encounter is defined as the first time two individuals come within 20m of one another during some span of time. Each node represents an individual, and each line represents an encounter that occurred during a particular quarter of one day. Quarter-days consist of the hours between 7am–11am, 11am–3pm, 3pm–7pm, and 7pm–11pm. There are twelve quarter-days across the three days of the study, and thus at most twelve lines that can connect any two nodes. Lines are coloured according to the quarter-day in which the encounter occurred, ranging from Thursday Quarter 1 (purple) to Saturday Quarter 4 (green). Node area is proportional to the total number of unique encounters for that individual across all three days (which would be that individual's degree, if the network were represented without temporal information).

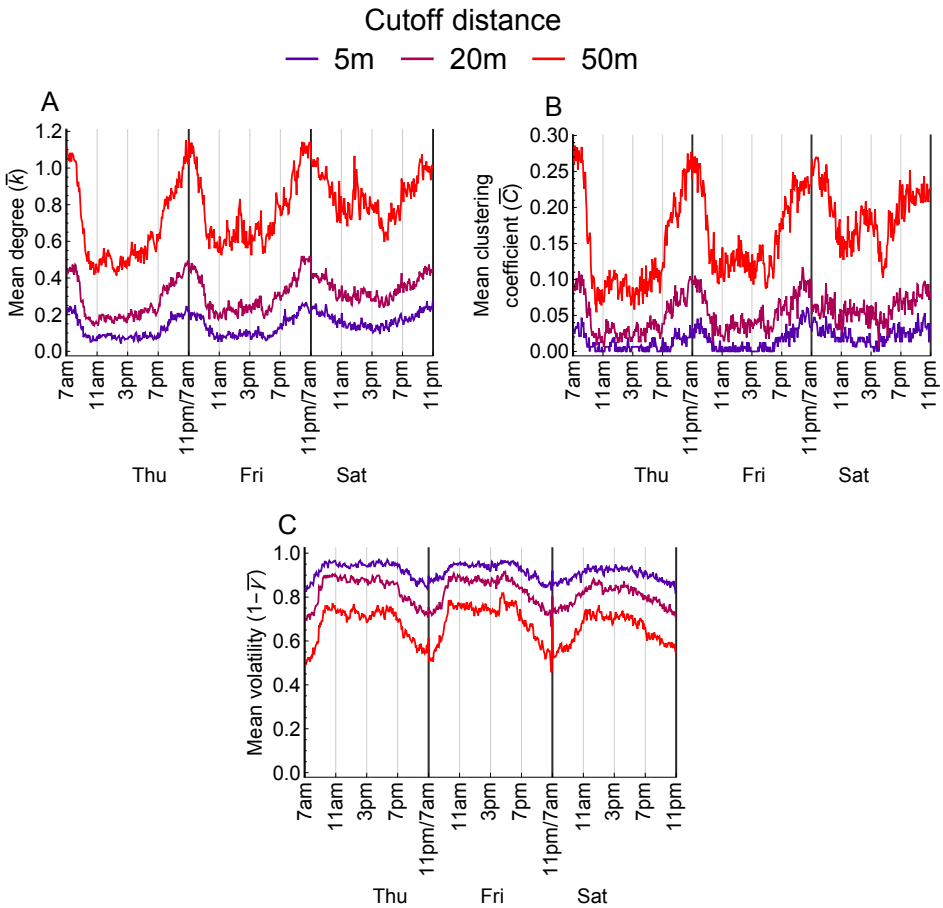

Figure 2: (A) Mean degree, (B) mean local clustering coefficient, and (C) mean link volatility over time for five-minute network 'snapshots' generated from the Haslemere dataset. Mean degree and mean local clustering coefficient are highest at night, while the mean link volatility is highest during the day.

3 The BBC Pandemic simulation

The simulated Haslemere outbreak is modelled as a susceptible-exposed-infectious (SEI) process for which the probability that an individual becomes infected at a given time is related to her/his distance from all currently infectious individuals. To fit into the programme narrative, the "outbreak" needed to be seeded by the presenter (user 469), and last for a total of three days. Because of this, some scientific liberties were taken: model parameter values were chosen to ensure that a substantial outbreak would occur within the three-day simulation window, requiring an unrealistically high basic reproduction number (see Section "Contrasting rival notions of R_0") and a short incubation period. Nevertheless, the model satisfies the central aim of communicating key aspects of respiratory infectious disease transmission, and can be easily adjusted to account for other scenarios. Full details on model specification are given in the Materials and Methods, and computer code for producing related simulations is given in Supplemental Code 1. All of the following references to the "Haslemere outbreak" refer to our computational simulations; no real illness was tracked in this study.

We used the transmission model to generate a single epidemic simulation, which we call the "the Haslemere outbreak", to feature in the BBC Four documentary. This was the very first run from the model using the full data; we did not run multiple simulations to choose one with "nice" properties. A total of 405 participants were 'infected' in the simulated Haslemere outbreak over the three days (Fig. 3A). The greatest number of infections occurred on Day 2, with the rate of infection increasing throughout Day 1 and declining throughout Day 3. A transmission tree for the outbreak is depicted in Fig. 3B. Each point in the tree represents an individual, and each line represents a direct infection. Each row in the tree therefore represents a new generation of transmission, where generations proceed from the top of the tree to the bottom. The longest transmission chain consists of 14 infections.

Identifying superspreaders

Individuals that are responsible for disproportionately many infections during an outbreak are commonly termed "superspreaders". In the virtual Haslemere outbreak, most (272) individuals caused no infections, but a few caused disproportionately many (Fig. 4A). The greatest number of direct

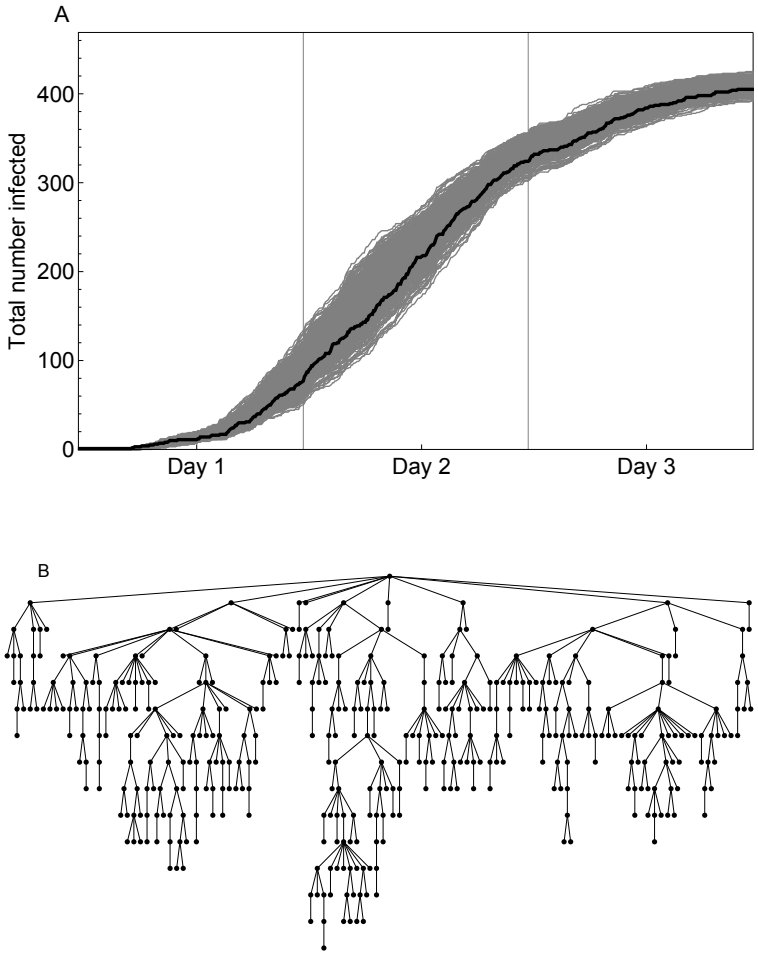

Figure 3: (A) Cumulative number of people infected over time in the Haslemere epidemic (black) and in 500 re-simulations (grey). By construction, the outbreak lasted for three days, after which 405 of the 469 participants were infected. The greatest number of infections happened on Day 2. All simulations are seeded with the same index case as the Haslemere epidemic. (B) Tree depicting direct infections in the Haslemere epidemic. Each row represents a new generation of the outbreak, from top to bottom. Note that infections that align on a row did not necessarily occur at the same time, but instead represent "generations" of the outbreak. The point at the root of the tree (top) represents the index case, who caused nine direct secondary infections. The longest chain consists of 14 infections.

secondary ('offspring') infections caused by a single individual was ten; interestingly, this person was not the virtual index case. Lloyd-Smith et al. [35] found that negative binomial distributions tend to fit a wide variety of offspring distributions observed during real outbreaks. The best-fit negative binomial distribution to the offspring distribution for the Haslemere outbreak has parameters $r = 0.640$ and $p = 0.426$ (also shown in Fig. 4A). The offspring distribution can be used to characterise superspreaders of infection in the following way: if $Z \sim NegBin(r, p)$ is the offspring distribution, an nth-percentile superspreader is anyone who infects at least $Z^{(n)}$ others, where $Z^{(n)}$ is the nth percentile of the offspring distribution [35]. For the Haslemere epidemic, the 90th-percentile superspreaders are those who infected at least $Z^{(90)} = 3$ individuals, of whom there were 47.

The superspreaders of the Haslemere outbreak were generally infected early in the epidemic, with all of the the 90th-percentile superspreaders infected before the afternoon of Day 2 (Fig. 4B). However, the converse is not true: not everyone who became infected early in the outbreak is a superspreader.

Individuals with many interpersonal encounters also tend to be superspreaders in the Haslemere epidemic, though again the converse is not true; that is, not all superspreaders have many encounters. Using the location data, we calculate the total number of encounters for each user, where an encounter is defined as the first time two users pass within 20 metres of each other. Fig. 4C provides a histogram of the number of unique encounters per user, distinguishing between the 90th-percentile superspreaders (red) and all others (grey). The users with the most encounters (above 41) are all superspreaders. However, the distribution of encounters made by the superspreaders is broad, with some superspreaders making as few as 10 encounters. On the other hand, some users with as many as 30 encounters are not superspreaders.

One might imagine that superspreaders could be predicted more accurately by a metric that takes into account the probability of infection. Using the transmission model (see Supplemental Materials and Methods) and the Haslemere mobility data, it is possible to calculate the expected number of people that a given participant in the Haslemere epidemic would directly infect if everyone else remained susceptible otherwise. We call this the user's 'individual reproduction number', denoted v_j for user j, following Lloyd-Smith et al. [35]. The quantity v_j is calculated as the probability $P_{i,j}$ that an infected individual j infects a susceptible individual i at some point over the

three days, summed over all individuals $i \neq j$ (full details given in the Supplemental Materials and Methods). The distribution of v_j for all users is depicted in Fig. 4D, again separated into the 90th-percentile superspreaders of the Haslemere epidemic (red) and all others (grey). As with the distribution of raw encounters, users with the highest v_j are superspreaders, but many superspreaders have low and even below-average v_j, suggesting that this too is an unreliable way of predicting which individuals will trigger the greatest number of secondary infections in a particular outbreak. Superspreaders are partly determined by the stochastic course of the outbreak itself and, in real outbreak settings, possibly by individual host-pathogen dynamics [39], and so superspreaders cannot in general be predicted accurately *a priori*.

4 Epidemiological relevance of dynamic interpersonal proximity data

The availability of dynamic interpersonal proximity data opens new avenues for investigating infectious disease outbreaks. Here, we use the Haslemere dataset to illustrate how dynamic population structure can influence empirical estimates of the basic reproduction number R_0 and may open new opportunities to identify optimal vaccination strategies. We use an extended version of the model that was used to generate the Haslemere epidemic, for which recovery is possible (thus becoming an SEIR model rather than an SEI model), random index cases are considered, and the data are temporally looped to allow for epidemics that last beyond three days. Full details are given in the Supplemental Materials and Methods.

Contrasting rival notions of R_0

The basic reproduction number (R_0) is a value of fundamental importance in infectious disease epidemiology. Defined as the expected number of secondary infections caused by a typical infectious individual introduced into a completely susceptible population [40], R_0 concisely quantifies the infectiousness of a disease in a given population, and is related to the expected final size of an outbreak and to the fraction of the population that would need to be vaccinated to prevent an outbreak from happening [41]. The mean individual reproduction number, $\bar{v} = 7.27$, gives a model-based estimate of the basic reproduction number for the Haslemere outbreak, in that it measures the expected number of secondary infections caused by a randomly chosen individual, if the rest

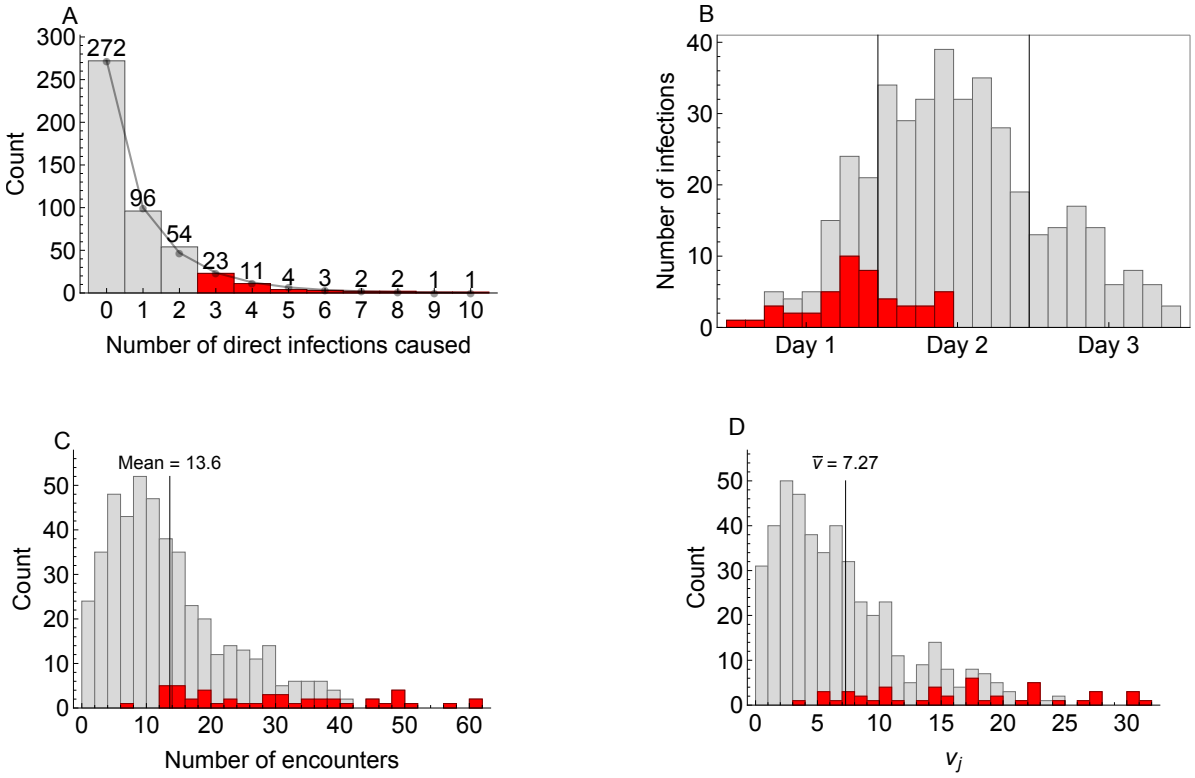

Figure 4: (A) Histogram of the number of direct secondary infections caused by each person in the Haslemere epidemic. The connected points depict the best-fit negative binomial distribution ($r = 0.640$, $p = 0.426$) to the histogram values. Over half (272) of the 469 participants caused no infections. The index case was responsible for directly infecting nine individuals. One person surpassed her, with ten direct infections. The outbreak's 90th-percentile superspreaders are the 10% of people ($n = 47$) who caused the greatest number of direct secondary infections (3 or more). These are highlighted in red. (B) Histogram of the total number of people infected over time (grey), and the number of the 90th-percentile superspreaders infected over time (red). All superspreaders are infected by the middle of Day 2 of the outbreak. (C) Histogram of the number of unique encounters within 20m for all users (grey) and for the 90th-percentile superspreaders (red). The mean number of unique encounters is 13.6. (D) Histogram of the individual reproduction number v_j for all users (grey) and for the 47 90th-percentile superspreaders (red). The mean individual reproduction number is $\bar{v} = 7.27$, which may also be interpreted as the overall basic reproduction number, R_0.

of the population were kept artificially susceptible [35].

Alternatively, there are many ways to empirically estimate R_0 from incidence data alone. An often-used method is based on the initial growth rate of the cumulative incidence during an outbreak [42]; another is based on the outbreak's final size [43] (see Supplemental Materials and Methods). Using outbreak simulations based on the Haslemere data, it is possible to see how these methods compare for realistic mixing patterns. Fig. 5 depicts the distribution of R_0 estimates from 1000 simulated outbreaks using the initial growth rate method (red) and the final size method (blue). Also depicted is the distribution of individual reproduction numbers v_j (black). The initial growth rate method in our case yields substantially higher and more variable R_0 estimates (mean: 14.0; IQR: 10.8–16.5) than the final size method (mean: 3.03; IQR: 2.97–3.10). The mean estimates of R_0 from the initial growth rate method and the final size method lie on either side of $\bar{v} = 7.27$. Both of the incidence-based R_0 estimation methods assume that the population is "well-mixed", or that each person has a uniform probability of coming into contact with every other. Their disagreement with each other and with the model-based estimate highlights the important point that deviations from the assumption of well-mixed contacts can profoundly affect empirical R_0 estimates. For populations with spatial or network-type structure, specifying R_0 can be challenging [44, 45], and the task is made even more complicated when that structure changes over time [46]. Accurately characterising R_0 for spatiotemporally-structured populations remains an important open problem that might be fruitfully explored using the Haslemere dataset.

Identifying optimal vaccination strategies

The Haslemere dataset provides a valuable opportunity to compare different vaccination strategies for realistic interpersonal mixing patterns. Targeted vaccination strategies, that preferentially vaccinate important transmitters of disease, should be more effective than random vaccination of the same number of individuals. It is not always clear how to define a given individual's importance for transmission, however: one might choose those with a diverse range of occasional contacts, those with many regular contacts, or those who act as 'bridges' between different sub-populations. Given complete movement data and a transmission model, the theoretically best way to identify an optimal vaccination set would be to design an algorithm that identifies the set of individuals

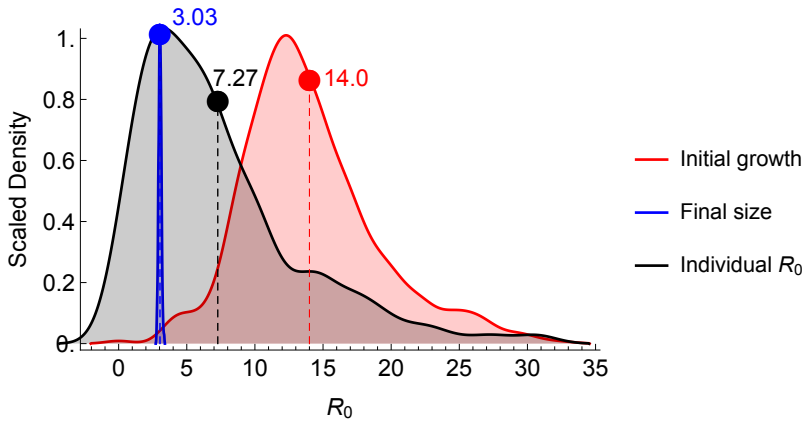

Figure 5: Distribution of R_0 as estimated using the initial growth formula (red) and the final size formula (blue) from 1,000 simulated outbreaks. Also depicted is the smoothed distribution of individual reproduction numbers, v_j (black; see also Fig. 4D). The coloured points depict the mean values of the corresponding distributions.

who, when vaccinated, tend to yield the smallest final epidemic size. A brute-force algorithm of this type is too computationally expensive to be feasible, however, so instead we here compare 13 different definitions of 'key transmitters', and use epidemic simulations to determine which definition yields the best vaccination set for minimising the final size of epidemics. These definitions are summarised in Table 1 and described in detail in the Supplemental Materials and Methods. Each vaccination set consists of 47 people, or 10% of the study population. The 13 targeted strategies are compared against three uniform-randomly chosen vaccination sets consisting of 10% of the population (labelled RV1–3), and against the scenario with no vaccination (labelled NV).

The distributions of final epidemic sizes under each vaccination scenario for 5,000 simulations are depicted in Fig. 6A. Vaccination clearly diminishes the final epidemic size, and the targeted vaccination strategies generally fare better than random vaccination. There is little variation in the final epidemic sizes among the targeted vaccination strategies, due in part to the long duration of infection in the simulated outbreaks. However, vaccination strategies RS10 and RSS appear to yield the greatest decrease in final epidemic size.

Sometimes, it is desirable not only to minimise the final size of an epidemic, but also to reduce but also to reduce its 'speed', or the rate of increase in incidence. In a real outbreak setting, this might buy time during which other interventions can be put into place. Fig. 6B depicts the distribution of times between the first and 234th infections for 5,000 simulated epidemics, for each

vaccination strategy. This represents the time it takes for 50% of the population to become infected. Again, the targeted vaccination strategies are 'better' than the random vaccination strategies, since the targeted strategies feature longer times-to-50%-infected. The strategies that minimise final size are not necessarily the ones that best slow the outbreak, however; the IR and IC strategies, for example, which are not especially effective at reducing final epidemic size compared with the other targeted strategies, are among the best at slowing the epidemic.

5 Discussion

This article presents a novel dataset that captures human interaction patterns at an unprecedented scope and level of spatiotemporal detail. Previous related datasets have been restricted to relatively controlled settings such as schools [12] and community gatherings [13], while the Haslemere dataset captures interactions for a subset of the population across an entire town for three consecutive days. Its temporal structure in particular makes it a valuable supplement to the contact surveys already used widely in epidemiological models [47, 48].

Though mobile phone geo-tracking capabilities have been leveraged to study human mobility before [3, 4, 5, 6, 7, 8, 9, 10], the spatial resolution in those studies is usually too coarse to characterise interpersonal interactions, and the data have generally not been made publicly available. Those studies have focused on characterising bulk movements, whereas the Haslemere dataset offers the novel opportunity to study fine-scale interactions between individuals as they occur across time.

The need to better understand the interplay between human mobility and disease transmission has long been recognised [49]. Proxies of human movement have been integrated into disease models with some success [50, 27, 51], but the advent of wearable sensing technology, like the GPS trackers included in most smartphones, is revolutionising our insight into the human-behavioural element of disease transmission. Mobility patterns affect not only the geographic spread patterns of disease [4], but also the intensity of outbreaks [52] and the genetic diversity of circulating pathogens [24]. While the characterisation of supposedly universal laws that govern human mobility has received considerable interest [53, 54, 10], it seems likely that a close attention to the rich variety of human movement patterns, across cultures, times, and geographic scales,

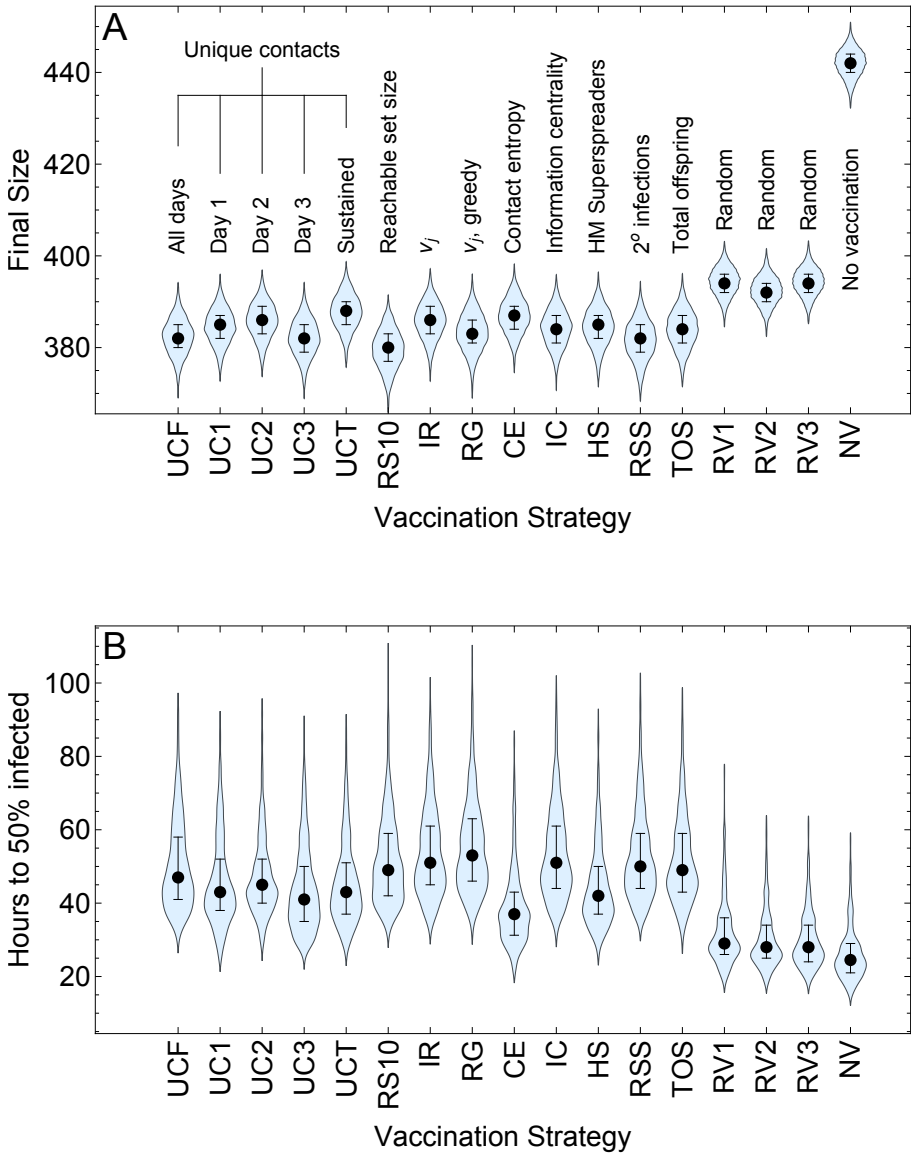

Figure 6: Violin plots depicting (A) the final epidemic size distributions and (B) the time required for 50% of the population to become infected, from 5,000 outbreak simulations for each of 17 vaccination strategies. The vaccination strategies are summarised in (A) and described fully in Table S1 and the Supplemental Information. The points depict the median value of the distribution, and the bars depict the 25% and 75% quartiles (see also Table S2). The first 13 vaccination strategies are summarised in Table 1. The final four strategies are controls, where strategies RV1-3 (for "random vaccination") each consist of vaccinating 47 randomly-chosen individuals, and strategy NV represents no vaccination. The final sizes and times-to-50%-infected for the epidemics that failed to establish (led to fewer than 10 total infections) are not depicted. These account for between 3% and 8% of the simulated outbreaks, depending on the vaccination strategy (see Table S2 for the exact values).

will yield the greatest new behavioural and epidemiological insights.

Our analysis of an epidemiological model based on the Haslemere dataset indicates that reliably predicting an outbreak's superspreaders and estimating its basic reproduction number remain challenging tasks. In time, the availability of detailed mobility data will help address these gaps, as well as to shed light on more fundamental questions, such as: what is the most epidemiologically relevant way to delineate a population? How can the spatiotemporal dynamics of human interactions best be captured mathematically? To what extent does *a priori* knowledge of an outbreak's superspreaders, R_0, and other key parameters facilitate its control? One of the clearest potential contributions of the Haslemere dataset and others like it will be the ability to evaluate intervention strategies under more realistic conditions. To illustrate this, we compare how thirteen different vaccination strategies fare under one particular transmission model built upon the Haslemere data. However, we do not take into account other types of interventions, nor any changes in behaviour that might accompany the spread of a disease; these will be important areas for future work. Different studies have found that social distancing during an outbreak can either dampen [55, 56] or exacerbate [57] transmission. Detailed mobility data may help explain under which conditions each scenario is likely, and thus help resolve this and other apparent contradictions in the field.

While the collection of human mobility data shows substantial promise for improving our understanding of disease dynamics, there remain major logistical and ethical challenges [58, 59]. The collection of the Haslemere dataset was made possible by a special collaboration between the media, academic scientists, and the broader public. The outreach event and study were made possible by the clout of the BBC and by the enthusiasm of the local press, museum, and public. Even under these arguably ideal circumstances, we were only able to collect three days' worth of data, which is too short of a duration to simulate a realistic influenza outbreak in a town. To maintain the anonymity of the volunteers, we aggregated the temporal data into five-minute bins and omitted nighttime observations. In the future, we recommend that similar studies collect data over longer time frames (on the order of weeks, if possible), that they restrict observations to every five minutes, and that they do not record nighttime movements. This should help strike a balance between relevance for infectious disease modelling and maintenance of the participants' privacy. We note that the use of voluntary data, like the Haslemere dataset, can be more ethically straightforward than that of so-called "convenience" datasets, but the collection of voluntary data also

requires far more planning and oversight. As further studies reveal the value of the Haslemere dataset, however, we hope that a strong case can be made for the collection of similar data across diverse geographic settings, and that this will contribute not just to epidemiology, but to human behavioural sciences as a whole.

6 Materials and Methods

Network measurements

The degree of a node is taken here to be the number of links that issue from it. The local clustering coefficient for a given node is the fraction of that node's neighbours that are also connected to one another with a link [37].

The link volatility of a node j, denoted $1 - \gamma_j$ by Clauset and Eagle (2012) [38], captures the correlation between elements of the adjacency matrices specified by two snapshots of a dynamic network. The correlation is only calculated between elements of the adjacency matrices that are nonzero for at least one of the two snapshots, to avoid spurious correlations between the zero elements in the matrices, which are often sparse. The link persistence is defined as

$$\gamma_j = \frac{\sum_i A_{i,j} A'_{i,j}}{\sqrt{(\sum_i A_{i,j})(\sum_i A'_{i,j})}} \tag{1}$$

where A and A' are (unweighted) adjacency matrices corresponding to two subsequent network snapshots. We define $\gamma_j = 0$ if node j has no neighhbours in either A or A'. Link persistence is bounded between 0 and 1. The volatility for link j is $1 - \gamma_j$.

The Haslemere epidemic model

For the Haslemere epidemic, individuals may progress from susceptible to exposed to infected (an SEI model). The exposure period lasts for 25 minutes (five time steps), after which the individual becomes infectious. There is no recovery; the epidemic is assumed to end at the end of the third day, when data collection concludes. At each time point, the force of infection $\lambda_{i,j}$ between

individuals i and j is modelled using an exponential curve with a cutoff:

$$\lambda_{i,j}(t) = \begin{cases} ae^{-d_{i,j}/\rho} & d_{i,j} \leq \xi \\ 0 & d_{i,j} > \xi. \end{cases} \tag{2}$$

Here, $d_{i,j}$ is the distance in metres between individuals i and j at time t, calculated from the users' cleaned location logs, a is the amplitude of the kernel, ρ is the 'characteristic distance' (the distance over which the kernel decreases by a factor of $1/e$), and ξ defines the cutoff distance, after which the force of infection is assumed to be zero. For the final model simulations, the parameter values are set at $a = 1$, $\rho = 10$ metres, and $\xi = 20$ metres (with associated kernel depicted in Fig. S1). The epidemic results were shared with the study participants in a public event immediately following the end of data collection, hence a fast decision on the model and parameters was necessary (within a few hours) and sensitivity analysis was somewhat limited. The values were chosen because they yield epidemic simulations that tend to expand over the course of a few days and infect a substantial fraction of the population. For the Haslemere outbreak, we identify the most likely infector as the geographically nearest infectious individual at the time of infection. This was for convenience; in subsequent simulations, such as when defining the vaccination strategies based on the number of secondary infections, we assign infectors randomly, with probability proportional to the force of infection contributed by each possible infector at the time of infection.

The probability of infection per unit time (5 minutes) is

$$P_i(t) = 1 - e^{-\lambda_i(t)} \tag{3}$$

following [60, 61].

The general epidemic model

For the investigation of estimation methods for R_0 and the analysis of vaccination strategies, we define a more general epidemic model that allows for recovery and loops through the three days of available data to allow for longer epidemics. This SEIR-type model uses Eq. 2 as the transmission kernel. Individuals recover and are immune to further infection after being infected for three days

(576 time steps).

Acknowledgements

We thank all those in Haslemere who took part in the *BBC Pandemic* study. We thank Adam Kucharski for his help in specifying the project, and for his assistance in developing the ideas presented in this article. We thank Hannah Fry for the interesting discussions regarding the final model. We thank 360 Production, especially Danielle Peck and Cressida Kinnear, for making possible the collection of the dataset that underlies this work.

References

[1] D. Peck, B. Cherrington, C. Kinnear, and 360 Production, "Contagion! The BBC Four Pandemic," 2018.

[2] P. Klepac, S. Kissler, and J. Gog, "Contagion! The BBC Four Pandemic - the model behind the documentary," *Epidemics*, 2018.

[3] J. R. B. Palmer, T. J. Espenshade, F. Bartumeus, C. Y. Chung, N. E. Ozgencil, and K. Li, "New Approaches to Human Mobility: Using Mobile Phones for Demographic Research," *Demography*, vol. 50, pp. 1105–1128, jun 2013.

[4] A. Wesolowski, N. Eagle, A. J. Tatem, D. L. Smith, A. M. Noor, R. W. Snow, C. O. Buckee, R. M. Prothero, P. Bejon, E. Dolgin, S. T. Stoddard, C. Lynch, C. Roper, A. J. Tatem, D. L. Smith, B. Moonen, A. L. Menach, R. Chuquiyauri, A. J. Tatem, C. Linard, M. Gilbert, R. W. Snow, A. M. Noor, A. J. Tatem, A. J. Tatem, A. M. Noor, C. von Hagen, A. D. Gregorio, S. I. Hay, A. J. Tatem, A. M. Noor, S. I. Hay, A. M. Noor, P. W. Gething, S. A. Mudhune, Y. Yé, N. Madise, R. Ndugwa, S. Ochola, R. W. Snow, S. Kasili, A. Wesolowski, N. Eagle, A. M. Noor, R. W. Snow, C. O. Buckee, M. C. González, C. A. Hidalgo, A. L. Barabási, C. Song, Z. Qu, N. Blumm, A. L. Barabási, C. Dye, G. Hasibeder, W. Sama, K. Dietz, T. Smith, D. Balcan, N. M. Ferguson, and I. M. Longini, "Quantifying the impact of human mobility on malaria.," *Science (New York, N.Y.)*, vol. 338, no. 6104, pp. 267–70, 2012.

[5] A. Wesolowski, T. Qureshi, M. F. Boni, P. R. Sundsøy, M. A. Johansson, S. B. Rasheed, K. Engø-Monsen, and C. O. Buckee, “Impact of human mobility on the emergence of dengue epidemics in Pakistan,” *Proceedings of the National Academy of Sciences*, vol. 112, no. 38, pp. 11887–11892, 2015.

[6] L. Bengtsson, J. Gaudart, X. Lu, S. Moore, E. Wetter, K. Sallah, S. Rebaudet, and R. Piarroux, “Using Mobile Phone Data to Predict the Spatial Spread of Cholera,” *Scientific Reports*, vol. 5, pp. 1–5, 2015.

[7] C. Song, Z. Qu, N. Blumm, and A.-L. Barabasi, “Limits of Predictability in Human Mobility,” *Science*, vol. 327, pp. 1018–1021, feb 2010.

[8] J. Reades, F. Calabrese, A. Sevtsuk, and C. Ratti, “Cellular Census: Explorations in Urban Data Collection,” *IEEE Pervasive Computing*, vol. 6, pp. 30–38, jul 2007.

[9] M. C. González, C. A. Hidalgo, and A.-L. Barabási, “Understanding individual human mobility patterns,” *Nature*, vol. 453, no. 7196, pp. 779–782, 2008.

[10] F. Simini, M. C. González, A. Maritan, and A.-L. Barabási, “A universal model for mobility and migration patterns.,” *Nature*, vol. 484, pp. 96–100, apr 2012.

[11] N. Eagle and A. (Sandy) Pentland, “Reality mining: sensing complex social systems,” *Personal and Ubiquitous Computing*, vol. 10, pp. 255–268, may 2006.

[12] H. Guclu, J. Read, C. J. Vukotich, D. D. Galloway, H. Gao, J. J. Rainey, A. Uzicanin, S. M. Zimmer, and D. A. Cummings, “Social contact networks and mixing among students in K-12 Schools in Pittsburgh, PA,” *PLoS ONE*, vol. 11, no. 3, pp. 1–19, 2016.

[13] C. Cattuto, W. van den Broeck, A. Barrat, V. Colizza, J. F. Pinton, and A. Vespignani, “Dynamics of person-to-person interactions from distributed RFID sensor networks,” *PLoS ONE*, vol. 5, no. 7, pp. 1–9, 2010.

[14] N. Eagle, A. S. Pentland, and D. Lazer, “Inferring friendship network structure by using mobile phone data,” *Proc. National Academy of Sciences of the United States of America (PNAS)*, vol. 106, no. 36, pp. 15274–15278, 2009.

[15] M. J. Keeling, L. Danon, M. C. Vernon, and T. A. House, “Individual identity and movement networks for disease metapopulations,” *Proceedings of the National Academy of Sciences*, vol. 107, pp. 8866–8870, may 2010.

[16] S. Riley, “Large-scale spatial-transmission models of infectious disease,” *Science*, vol. 316, no. 5829, pp. 1298–1301, 2007.

[17] V. Colizza, M. Barthélemy, A. Barrat, and A. Vespignani, “Epidemic modeling in complex realities,” *Comptes Rendus Biologies*, vol. 330, pp. 364–374, apr 2007.

[18] M. J. Keeling and K. T. Eames, “Networks and epidemic models,” *Journal of The Royal Society Interface*, vol. 2, no. 4, pp. 295–307, 2005.

[19] L. Danon, A. P. Ford, T. House, C. P. Jewell, M. J. Keeling, G. O. Roberts, J. V. Ross, and M. C. Vernon, “Networks and the Epidemiology of Infectious Disease,” *Interdisciplinary Perspectives on Infectious Diseases*, vol. 2011, pp. 1–28, 2011.

[20] F. Liljeros, C. R. Edling, and L. a. N. Amaral, “Sexual networks: Implication for the transmission of sexually transmitted infection,” *Microbes and Infections*, vol. 5, pp. 189–196, 2003.

[21] L. Pellis, F. Ball, S. Bansal, K. Eames, T. House, V. Isham, and P. Trapman, “Eight challenges for network epidemic models,” *Epidemics*, vol. 10, pp. 58–62, 2015.

[22] S. Bansal, B. T. Grenfell, and L. A. Meyers, “When individual behaviour matters: homogeneous and network models in epidemiology,” *Journal of The Royal Society Interface*, vol. 4, no. 16, pp. 879–891, 2007.

[23] J. M. Read and M. J. Keeling, “Disease evolution on networks: the role of contact structure,” *Proceedings of the Royal Society B: Biological Sciences*, vol. 270, no. 1516, pp. 699–708, 2003.

[24] C. O. Buckee, K. Koelle, M. J. Mustard, and S. Gupta, “The effects of host contact network structure on pathogen diversity and strain structure,” *Proceedings of the National Academy of Sciences*, vol. 101, no. 29, pp. 10839–10844, 2004.

[25] M. S. Handcock and J. H. Jones, “Likelihood-based inference for stochastic models of sexual network formation,” *Theoretical Population Biology*, vol. 65, no. 4, pp. 413–422, 2004.

[26] S. M. Firestone, M. P. Ward, R. M. Christley, and N. K. Dhand, “The importance of location in contact networks: Describing early epidemic spread using spatial social network analysis,” *Preventive Veterinary Medicine*, vol. 102, no. 3, pp. 185–195, 2011.

[27] D. Balcan, V. Colizza, B. Goncalves, H. Hu, J. J. Ramasco, and A. Vespignani, “Multiscale mobility networks and the spatial spreading of infectious diseases,” *Proceedings of the National Academy of Sciences*, vol. 106, pp. 21484–21489, dec 2009.

[28] J. Enright and R. R. Kao, “Epidemics on dynamic networks,” *Epidemics*, vol. 24, no. July 2016, pp. 88–97, 2018.

[29] S. Bansal, J. Read, B. Pourbohloul, and L. A. Meyers, “The dynamic nature of contact networks in infectious disease epidemiology,” *Journal of Biological Dynamics*, vol. 4, no. 5, pp. 478–489, 2010.

[30] N. Masuda and P. Holme, “Predicting and controlling infectious disease epidemics using temporal networks,” *F1000Prime Reports*, vol. 5, no. March, 2013.

[31] M. Morris, “Network Epidemiology: A Handbook for Survey Design and Data Collection,” *Network Epidemiology: A Handbook for Survey Design and Data Collection*, no. April, pp. 1–252, 2004.

[32] M. Morris, “Network Dynamism: History and Lessons of the Colorado Springs Study,” in *Network Epidemiology*, pp. 87–114, Oxford University Press, mar 2004.

[33] M. C. Vernon and M. J. Keeling, “Representing the UK’s cattle herd as static and dynamic networks,” *Proceedings of the Royal Society B: Biological Sciences*, vol. 276, no. 1656, pp. 469–476, 2009.

[34] J. A. Drewe, K. T. D. Eames, J. R. Madden, and G. P. Pearce, “Integrating contact network structure into tuberculosis epidemiology in meerkats in South Africa: Implications for control,” *Preventive Veterinary Medicine*, vol. 101, no. 1-2, pp. 113–120, 2011.

[35] J. O. Lloyd-Smith, S. J. Schreiber, P. E. Kopp, and W. M. Getz, “Superspreading and the effect of individual variation on disease emergence.,” *Nature*, vol. 438, no. 7066, pp. 355–359, 2005.

[36] K. Ognyanova, “Static and dynamic network visualization with R,” 2017.

[37] D. J. Watts and S. H. Strogatz, “Collective dynamics of ’small-world’ networks.,” *Nature*, vol. 393, no. 6684, pp. 440–2, 1998.

[38] A. Clauset and N. Eagle, “Persistence and periodicity in a dynamic proximity network,” pp. 1–5, 2012.

[39] R. A. Stein, “Super-spreaders in infectious diseases,” *International Journal of Infectious Diseases*, vol. 15, no. 8, pp. e510–e513, 2011.

[40] O. Diekmann, J. A. P. Heesterbeek, and J. A. J. Metz, “On the definition and the computation of the basic reproduction ratio R0 in models for infectious diseases in heterogeneous populations.,” *Journal of Mathematical Biology*, vol. 28, no. 4, pp. 365–382, 1990.

[41] G. Chowell, H. Nishiura, and L. M. Bettencourt, “Comparative estimation of the reproduction number for pandemic influenza from daily case notification data,” *Journal of The Royal Society Interface*, vol. 4, pp. 155–166, feb 2007.

[42] M. G. Roberts and J. A. P. Heesterbeek, “Model-consistent estimation of the basic reproduction number from the incidence of an emerging infection,” *Journal of Mathematical Biology*, vol. 55, pp. 803–816, nov 2007.

[43] E. Vynnycky and R. G. White, *An introduction to infectious disease modelling*. Oxford: Oxord University Press, 2010.

[44] S. Riley, K. Eames, V. Isham, D. Mollison, and P. Trapman, “Five challenges for spatial epidemic models,” *Epidemics*, vol. 10, pp. 68–71, 2015.

[45] R. Breban, R. Vardavas, and S. Blower, “Theory versus data: How to calculate R0?,” *PLoS ONE*, vol. 2, no. 3, pp. 0–3, 2007.

[46] P. Holme and N. Masuda, “The basic reproduction number as a predictor for epidemic outbreaks in temporal networks,” *PLoS ONE*, vol. 10, no. 3, pp. 1–15, 2015.

[47] J. Mossong, N. Hens, M. Jit, P. Beutels, K. Auranen, R. Mikolajczyk, M. Massari, S. Salmaso, G. S. Tomba, J. Wallinga, J. Heijne, M. Sadkowska-Todys, M. Rosinska, and W. J. Edmunds, "Social Contacts and Mixing Patterns Relevant to the Spread of Infectious Diseases," *PLoS Medicine*, vol. 5, p. e74, mar 2008.

[48] O. le Polain de Waroux, S. Flasche, A. J. Kucharski, C. Langendorf, D. Ndazima, J. Mwanga-Amumpaire, R. F. Grais, S. Cohuet, and W. J. Edmunds, "Identifying human encounters that shape the transmission of Streptococcus pneumoniae and other acute respiratory infections," *Epidemics*, 2018.

[49] R. M. Prothero, "Disease and mobility: A neglected factor in epidemiology," *International Journal of Epidemiology*, vol. 6, no. 3, pp. 259–267, 1977.

[50] M. Tizzoni, P. Bajardi, A. Decuyper, G. Kon Kam King, C. M. Schneider, V. Blondel, Z. Smoreda, M. C. González, and V. Colizza, "On the Use of Human Mobility Proxies for Modeling Epidemics," *PLoS Computational Biology*, vol. 10, no. 7, 2014.

[51] C. Peng, X. Jin, K.-C. Wong, M. Shi, and P. Liò, "Collective Human Mobility Pattern from Taxi Trips in Urban Area," *PLoS ONE*, vol. 7, p. e34487, apr 2012.

[52] B. D. Dalziel, B. Pourbohloul, and S. P. Ellner, "Human mobility patterns predict divergent epidemic dynamics among cities.," *Proceedings. Biological sciences / The Royal Society*, vol. 280, no. 1766, p. 20130763, 2013.

[53] A. Noulas, S. Scellato, R. Lambiotte, M. Pontil, and C. Mascolo, "A tale of many cities: Universal patterns in human urban mobility," *PLoS ONE*, vol. 7, no. 5, 2012.

[54] D. Brockmann, L. Hufnagel, and T. Geisel, "The scaling laws of human travel.," *Nature*, vol. 439, no. 7075, pp. 462–5, 2006.

[55] R. Glass, L. Glass, W. Beyeler, and H. Min, "Targeted Social Distancing Designs for Pandemic Influenza," *Emerging Infectious Diseases*, vol. 12, no. 11, pp. 1671–1681, 2006.

[56] S. Kumar, J. J. Grefenstette, D. Galloway, S. M. Albert, and D. S. Burke, "Policies to reduce influenza in the workplace: Impact assessments using an agent-based model," *American Journal of Public Health*, vol. 103, no. 8, pp. 1406–1411, 2013. PMID: 23763426.

[57] S. Meloni, N. Perra, A. Arenas, S. Gómez, Y. Moreno, and A. Vespignani, “Modeling human mobility responses to the large-scale spreading of infectious diseases,” *Scientific Reports*, vol. 1, pp. 1–7, 2011.

[58] Y. A. De Montjoye, C. A. Hidalgo, M. Verleysen, and V. D. Blondel, “Unique in the Crowd: The privacy bounds of human mobility,” *Scientific Reports*, vol. 3, pp. 1–5, 2013.

[59] C. O. Buckee, “Protect privacy of mobile data,” *Nature*, vol. 514, pp. 35–35, oct 2014.

[60] R. M. Eggo, S. Cauchemez, and N. M. Ferguson, “Spatial dynamics of the 1918 influenza pandemic in England, Wales and the United States,” *Journal of the Royal Society, Interface*, vol. 8, pp. 233–43, feb 2011.

[61] J. R. Gog, S. Ballesteros, C. Viboud, L. Simonsen, O. N. Bjørnstad, J. Shaman, D. L. Chao, F. Khan, and B. T. Grenfell, “Spatial Transmission of 2009 Pandemic Influenza in the US,” *PLoS Computational Biology*, vol. 10, p. e1003635, jun 2014.