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*

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

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 for that documentary 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 interpersonal distance dataset. We discuss salient features of the dataset, including daily patterns in the clustering and volatility of interpersonal interactions. Our epidemiological analysis of the simulated Haslemere outbreak serves as a springboard to discuss scientific opportunities opened by the Haslemere dataset and others like it. 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 Influenza 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 continues on to infect the rest of the UK. In this scientific companion to the documentary, we present a novel dataset that captures the interpersonal distances between 469 volunteers from Haslemere over three days, and we describe a mathematical model that incorporates these data, used to simulate the “Haslemere outbreak” featured in the television programme. A description of the UK-wide model is given by Klepac *et al.* (2018) [2].

The production of the documentary provided an ideal opportunity to engage the public in collecting scientifically valuable data on interpersonal contact patterns. Close-proximity interactions are the main avenue by which many infections spread, and yet we lack a detailed understanding of when, with whom, and with what frequencies and durations these interactions normally occur. This is due in part to a lack of adequate data, which can in turn be traced to technological limitations and to issues with privacy. Wearable proximity sensors can provide detailed measurements of interpersonal distances [3, 4], but these devices can be expensive and require substantial oversight in their deployment and collection, limiting the scope of studies that rely on them. Mobile phones are an attractive alternative, since many people carry them and they feature a range of sensing technologies that can both identify nearby devices and measure GPS location. Bluetooth, for example, can identify nearby devices with a simple ‘is/is not within signal range’ dichotomy [5, 6], but is unreliable for measuring actual distances. Mobile phone GPS capabilities, on the other hand, can pinpoint a user’s location using various strategies including satellite-mediated GPS, cellular tower triangulation, and wifi connection. These GPS tags can then be used to measure the distances between phones. GPS accuracy is imperfect [7], however, and the interpersonal distances measured by GPS cannot take physical barriers (e.g. walls) between the users into account. A range of studies have used mobile phones to study human mobility patterns [3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14],

but due in part to privacy concerns the majority of these only consider bulk movements, obscuring the fine-scale interaction patterns that underpin the spread of disease. The study by Palmer *et al.* (2013) [7] most closely resembles the one presented here, though the scope is global rather than within a single town, covers fewer participants, and the dataset is not publicly available.

Our understanding of the nature of interpersonal interactions is also limited by a lack of adequate mathematical vocabulary to describe their intricate, often 'fleeting' structure [7, 15]. Networks are currently a popular way of incorporating population structure into models of infectious disease transmission [16, 17, 18, 19, 20, 21, 22, 23]. Static, unweighted networks are particularly amenable to mathematical analysis, and have featured in a range of disease transmission models [22, 24, 23, 25, 26, 27, 28]. However, there is a growing body of evidence that suggests that static networks 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 [15, 29, 30] and empirical [31, 32, 33, 34] standpoints. Datasets like the one presented here may help ground future dynamic network models in reality, and may help reveal what information is lost when the the rich dynamics of human interactions are approximated by a static network.

The Haslemere dataset enhances our capability to examine disease transmission and human behaviour in general. While we focus primarily on the work carried out for the BBC Four documentary, we also briefly discuss 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 we examine the challenges associated with empirically estimating the basic reproduction number R_0 for outbreaks in structured populations. We conclude with a discussion of lessons learned and open questions that might be addressed using 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. The volunteers constitute a convenience sample of 4.2% of the total population of Haslemere, according to the 2011 UK census [36]. Participants downloaded the *BBC Pandemic* mobile phone app and then went about their daily business, with the app running in the background. Ethical guidelines required that the study be restricted to volunteers at least 16 years of age, or at least 13 years of age with parental consent. The pairwise distances between volunteers were calculated using Haversine formula for great-circle geographic distance and are based on the most accurate GPS coordinates that the volunteers' mobile phones could provide. Full details on the study site, recruitment, data collection, and post-processing are provided in the SI.

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 [37]. This illustrates the heterogeneity and temporal complexity of the data, 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 SI), at four different temporal resolutions. The video illustrates how some connections persist over time, while other dynamic elements are lost when the network is temporally integrated.

Although the Haslemere dataset is among the most comprehensive of its kind, it is still subject to a number of limitations. Since a substantial fraction of Haslemere residents did not participate in the study, the dataset (and any network representation, such as Fig. 1) is unrealistically sparse. The interaction patterns are also biased by participant age, since children under 13 were not allowed to participate and the rate of smartphone ownership declines substantially in adults over 65 [38]. A literal interpretation of the contact network is therefore inadvisable; instead, the dataset should be understood to represent a proportion of adult encounters, from which general patterns may be identified with some caution. Epidemic models based on the Haslemere dataset should take these limitations into account.

Interaction patterns by time of day

The Haslemere dataset provides an opportunity to examine how interpersonal interaction patterns within a town vary by time of day. One might expect individuals to have many fleeting interactions

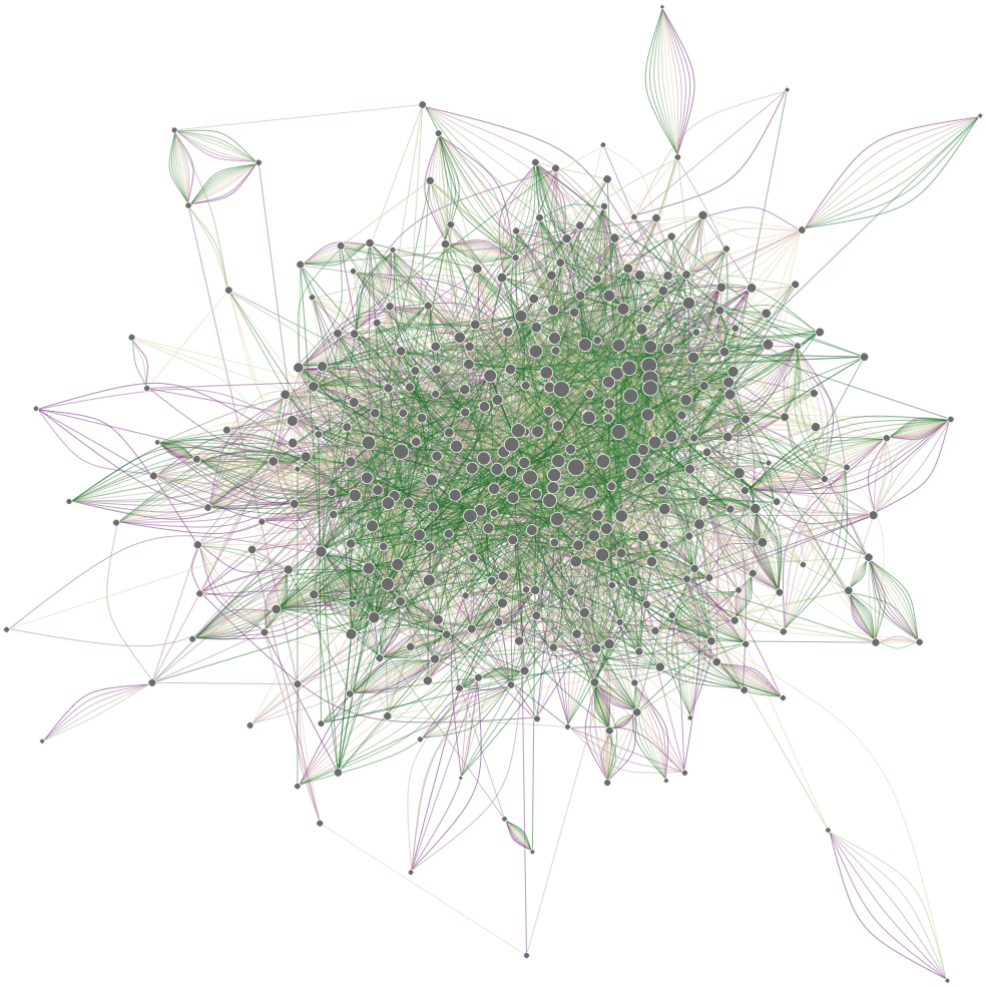

Figure 1: Network depicting pairwise encounters over time for the 75% of individuals in the Haslemere dataset who spend the greatest amount of time within 20m of another person (i.e. with the most total 'person-hours' of contact). An encounter is defined as the first time two individuals come within 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).

during the day, and fewer, more persistent interactions at night. The Haslemere dataset supports this hypothesis. To illustrate this, we calculate the mean degree, mean local clustering coefficient [39], and mean link volatility [40] (Fig. 2) during five-minute network 'snapshots' generated from the pairwise Haslemere 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) [40], extends the Pearson correlation coefficient to measure the persistence (γ) of network links between subsequent snapshots (see Materials and Methods). For all cutoff distances, 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). This suggests 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 [40], 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 the Haslemere study captures the movements of people across a whole town, and thus represents a more general cross-section of the population.

These findings suggest that interpersonal interaction patterns might productively be separated into a 'daytime' and 'nighttime' regime, or possibly into 'school/work' and 'home'. For epidemiological applications, one might be most interested in daytime interactions when describing the spread of a disease transmitted through casual contact (e.g. measles), while one might wish to know more about nighttime contacts when describing the spread of a disease transmitted mainly through persistent contact (e.g. tuberculosis). For still other diseases, some combination of the two will be relevant. We recommend including time of day and/or geographic venue as an element in future interpersonal contact surveys that are intended to support epidemiological research.

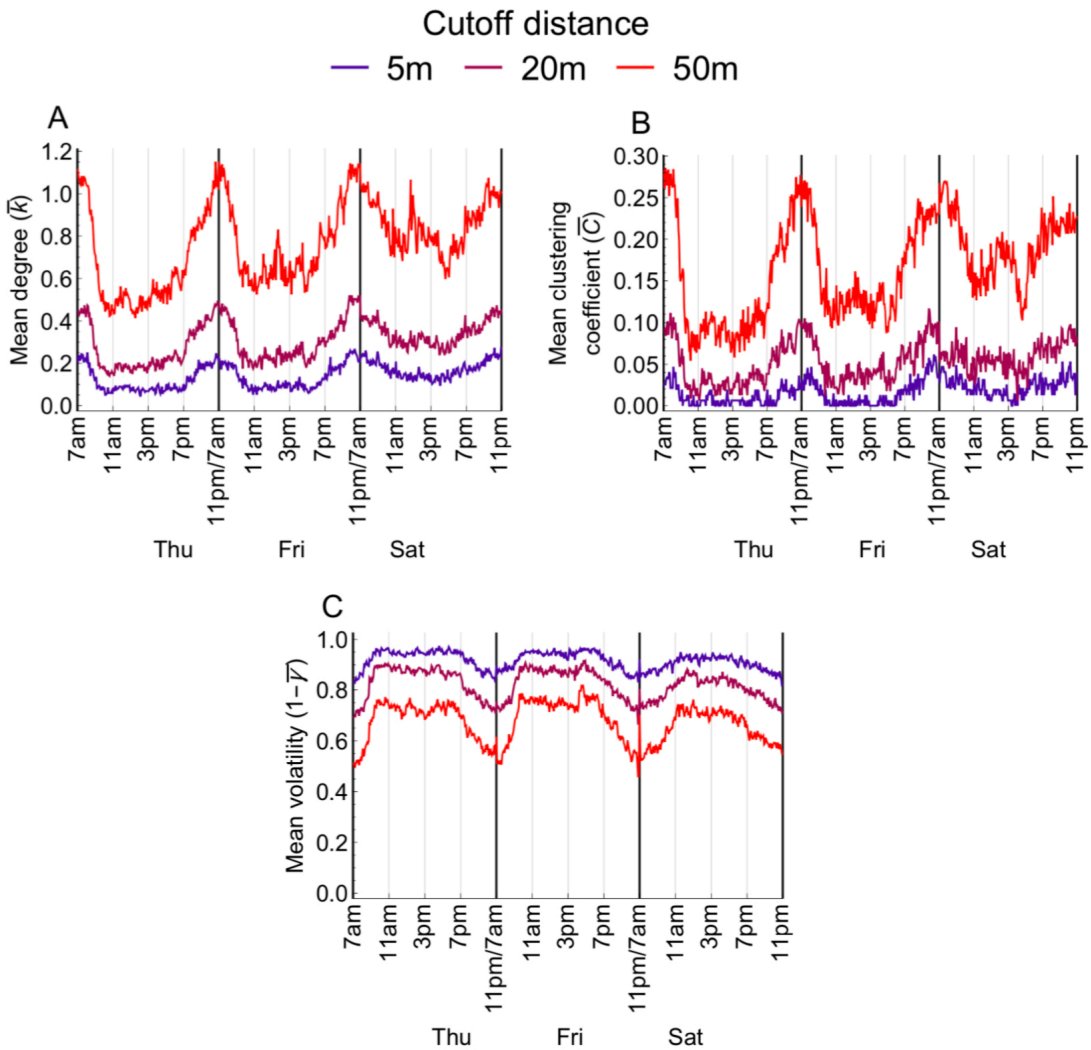

Figure 2: (A) Mean degree, (B) mean local clustering coefficient, and (C) mean link volatility over time for five-minute network 'snapshots' generated from the Haslemere dataset. Mean degree and mean local clustering coefficient are highest at night, while the mean link volatility is highest during the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

3 The BBC Pandemic simulation

To generate the simulated outbreak for the BBC documentary, we constructed a minimal susceptible-exposed-infectious (SEI) model that incorporates the Haslemere dataset. Here, we discuss that model and present a brief epidemiological analysis. The Haslemere outbreak simulation was produced in real time during a three-day data collection period. The participants' GPS locations were sent nightly to the authors of this article to develop and test a disease transmission model. Immediately following the final data pull, we simulated a single outbreak from the model to feature in the documentary. 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 (R_0) and a short incubation period. Motivation for the chosen model parameter values may be found in the SI. Despite these liberties, this minimal 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. A total of 405 participants were 'infected' in the simulated Haslemere outbreak over the three days (Fig. 3A). Generations of infection were identified by assigning to each infection an immediate 'parent', taken to be the geographically nearest infectious individual at the time of infection. While this stretches the underlying theory somewhat – infections are actually caused by the accumulation of forces from nearby infectious individuals, not by a single progenitor – this approach serves to illustrate key transmission routes for this particular simulation. Fig. 3B depicts these generations of infection as a transmission tree. The longest transmission chain consists of 14 infections.

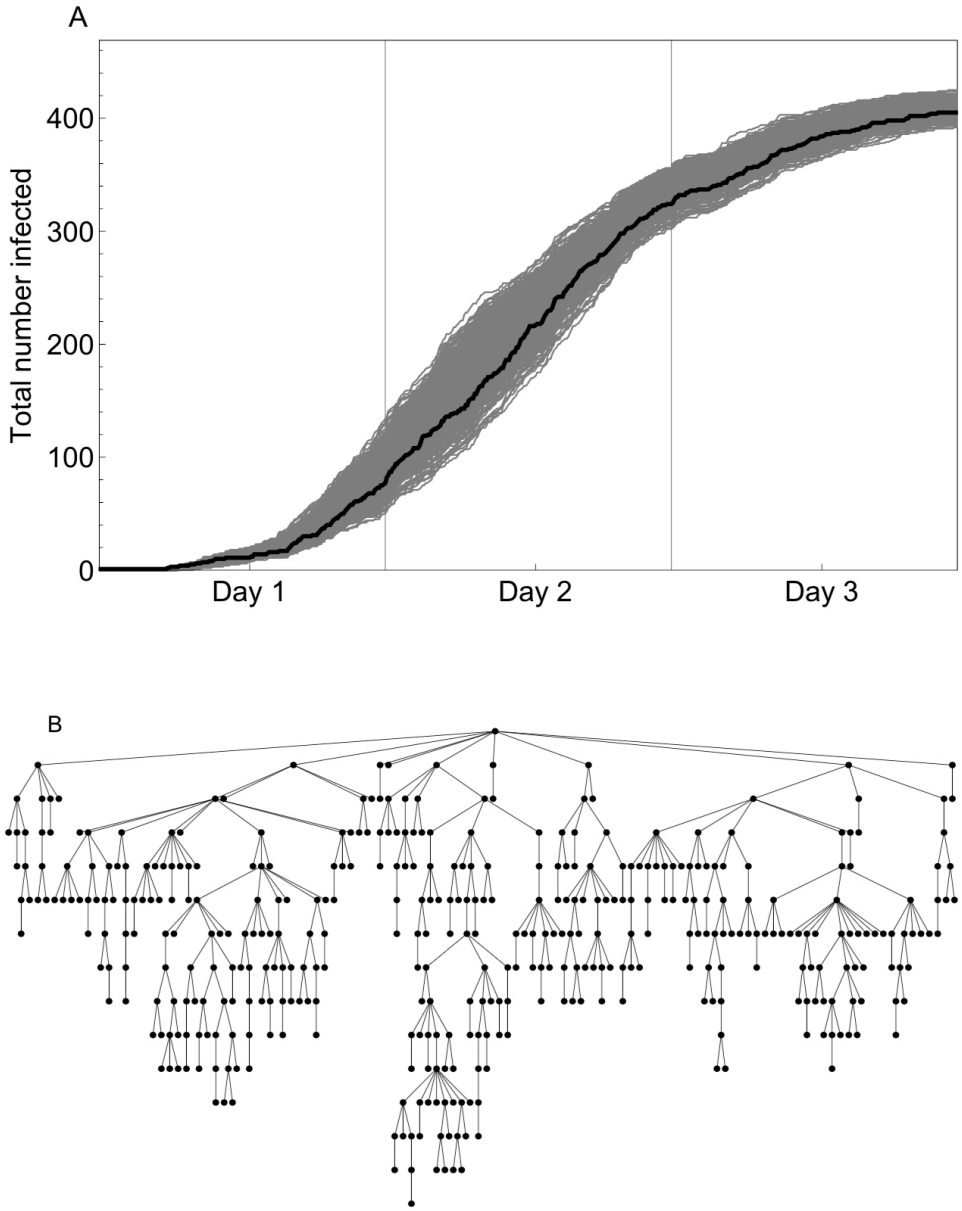

Figure 3: (A) Cumulative number of people infected over time in the Haslemere epidemic (black) and in 500 re-simulations (grey). (B) Tree depicting direct infections in the Haslemere epidemic. Each row represents a new generation of the outbreak, from top to bottom. Note that infections that align on 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.

Identifying superspreaders

Individuals that are responsible for disproportionately many infections during an outbreak are commonly termed "superspreaders". Superspreaders are natural targets for interventions aimed at reducing disease transmission, but are notoriously difficult to identify [35, 41, 42]. Here, we ask whether the superspreaders of the Haslemere outbreak can be identified using time of infection (are the superspreaders simply those who happened to get infected early in an outbreak?), number of encounters (are the superspreaders those who interact with the greatest number of people?), or a model-based individual reproduction number (are the superspreaders those with the greatest infectious potential?). To identify the superspreaders of the Haslemere outbreak, we use our earlier strategy of tracing each infection to a "most likely" infectious progenitor. By this metric, most (272) individuals caused no infections, but a few caused disproportionately many (Fig. 4A). Lloyd-Smith et al. [35] found that negative binomial distributions tend to fit a wide variety of offspring distributions observed during real outbreaks. If $Z \sim NegBin(r, p)$ is the offspring distribution, an nth-percentile superspreader is defined as 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 (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 SI) 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 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 superspreaders. Superspreaders are partly determined by the stochastic course of the outbreak itself and, in real outbreak settings, possibly by individual host-pathogen dynamics [42], and so superspreaders cannot in general be predicted accurately *a priori*.

Comparing rival notions of R_0

Next, we use the Haslemere dataset to illustrate how dynamic population structure can influence empirical estimates of the basic reproduction number R_0. 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 SI.

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 [43], R_0 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 [44]. 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 of the population were kept artificially susceptible [35].

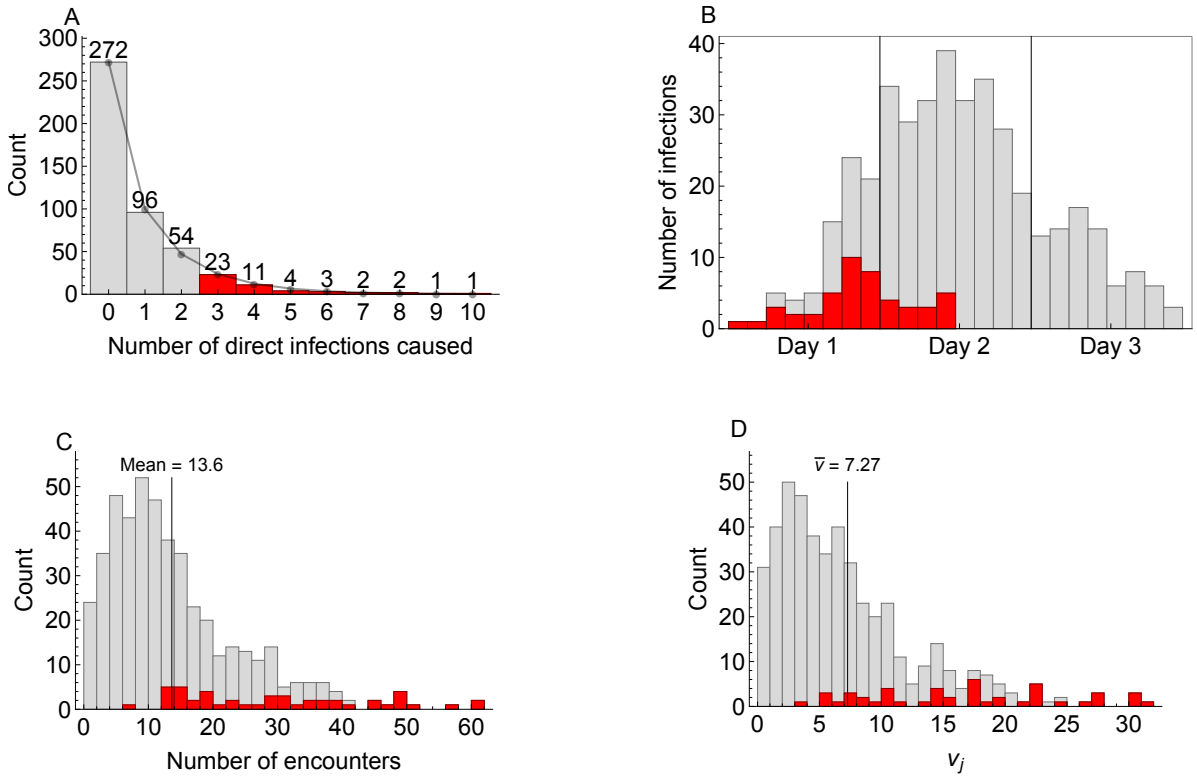

Figure 4: (A) Histogram of the number of direct secondary infections caused by each person in the Haslemere epidemic. The connected points depict the best-fit negative binomial distribution ($r = 0.640$, $p = 0.426$) to the histogram values. The outbreak's 90th-percentile superspreaders are the 10% of 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.

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 [45]; another is based on the outbreak's final size [46] (see SI). Both of these empirical methods assume that the underlying population is "well-mixed", which clearly does not hold for the population of Haslemere. We ask whether these methods can nevertheless provide accurate estimates of R_0.

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). Though the mean values of the distributions differ somewhat due to their spread (mean initial growth $R_0 = 4.0$; mean final size $R_0 = 3.0$; $\bar{v} = 7.3$), the modes of the distributions align remarkably well. However, accurate initial growth and final size R_0 estimates rely on accurate knowledge of the total population size and the disease's generation interval, respectively. For the Haslemere simulation, the size of the population is known exactly, and the generation interval may be measured directly using the span of time between each infection and its 'parent' (here, we set the generation interval to be the median of these time spans). In a real outbreak setting, these values may be uncertain, profoundly affecting estimates of R_0. For populations with spatial or network-type structure, specifying R_0 is challenging [47, 48], and the task is made even more complicated when that structure changes over time [49]. The Haslemere dataset and others like it may contribute to these discussions by allowing researchers to more precisely quantify how uncertainty in fundamental parameters propagates to estimates of R_0 in the context of realistic population structures.

4 Discussion

This article presents a novel open-source 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 [3] and community gatherings [4] or to a particular sub-population (e.g. students [40, 50, 51]), while the Haslemere dataset captures interactions for a subset of the population across an entire town for three consecutive days. The Haslemere dataset's temporal structure in particular makes it a valuable supplement to the con-

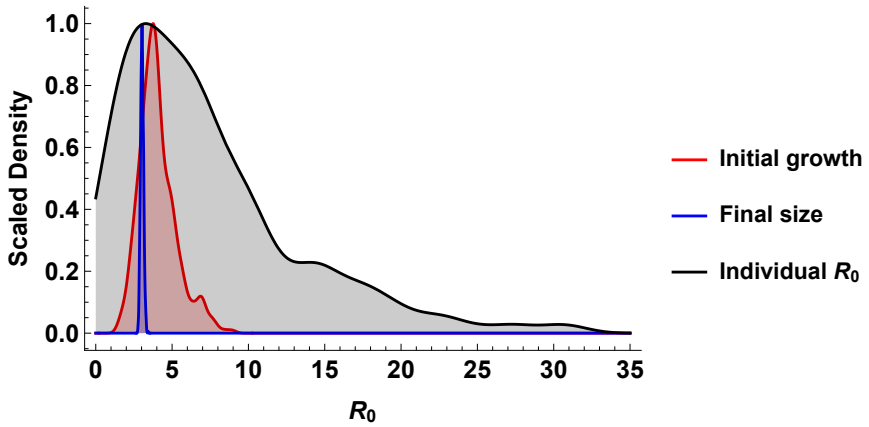

Figure 5: Distribution of R_0 as estimated using the initial growth formula (red, mean = 4.0) and the final size formula (blue, mean = 3.0) from 1,000 simulated outbreaks. Also depicted is the smoothed distribution of individual reproduction numbers, v_j (black, mean = 7.3; see also Fig. 4D).

tact surveys already used widely in epidemiological models [52, 53]. The Haslemere dataset is limited in that it does not include the interaction patterns of children under the age of 13, who are frequently implicated as important transmitters of respiratory disease [54, 55, 56]. Using GPS to measure interpersonal distances also means that physical barriers that could obstruct disease transmission are not accounted for. We believe these limitations are a reasonable trade-off for a dataset that is freely available (precluding the inclusion of data from children) and that provides distance measurements on a continuum while minimising the barrier to participation (using mobile phones, rather than custom proximity sensors).

Though mobile phone geo-tracking capabilities have been leveraged to study human mobility before [7, 8, 9, 10, 11, 12, 13, 14], the spatial resolution in those studies is usually too coarse to characterise interpersonal interactions, and/or the data have generally not been made publicly available. Most of these studies have focused on characterising bulk movements, whereas the Haslemere dataset enhances our ability 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 [57]. Proxies of human movement have been integrated into disease models with some success [58, 28, 59], 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 [8], but also the intensity of outbreaks [60] and the genetic diversity of circulating pathogens [25]. While the characterisation of supposedly universal laws that govern human mobility has received considerable interest [61, 62, 14], it seems likely that a close attention to the rich variety of human movement patterns, across cultures, times, and geographic scales, 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. Importantly, however, the Haslemere dataset does not account for behaviour changes that might occur in the context of a real outbreak. Different studies have found that social distancing during an outbreak can either dampen [63, 64] or exacerbate [65] transmission, underscoring the need for further investigation.

While the collection of human mobility data shows substantial promise for improving our understanding of disease dynamics, there remain major logistical and ethical challenges [66, 67]. The collection of the Haslemere dataset was made possible by a special collaboration between the media, academic scientists, and the broader public. The study was 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. The dataset is also

limited in that the GPS-mediated location tags from the phone's operating system do not capture epidemiologically relevant physical barriers between individuals, such as walls. This could be surmounted for example by using wearable wireless proximity sensors as in [50], but at the significant cost of requiring volunteers to use a piece of technology not already integrated into their everyday lives. Distributing, maintaining, and recovering the requisite sensors would raise the barrier for participation, likely reducing the number of volunteers. We note that the collection of voluntary data, like the Haslemere dataset, requires more planning and oversight than the collection of so-called "convenience" datasets used in some previous mobile phone-based studies [10, 4, 58]), but it has the benefit of being more ethically straightforward, since all participants give explicit approval for the use of their data. As further studies reveal the value of the Haslemere dataset, 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.

5 Materials and Methods

Network measurements

The degree of a node is 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 [39].

The volatility of a node j, denoted $1 - \gamma_j$ by Clauset and Eagle (2012) [40], is related to the correlation between 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 node 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 real distance in metres between individuals i and j at time t, a is the amplitude of the kernel, ρ is the decay rate, and ξ is the cutoff distance, after which the force of infection drops to zero. For the simulations, the parameter values are $a = 1$, $\rho = 10$ metres, and $\xi = 20$ metres (with associated kernel depicted in Fig. S1). The simulation 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 limited. 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 [68, 69].

The general epidemic model

To investigate R_0 estimation methods and 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] 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.

[4] 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.

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

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

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

[12] 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.

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

[14] 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.

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

[16] 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.

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

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

[19] 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.

[20] 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.

[21] 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.

[22] 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.

[23] 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.

[24] 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.

[25] 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.

[26] 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.

[27] 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.

[28] 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.

[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] UK Office for National Statistics, "2011 Census summaries for towns and villages in Waverly," tech. rep., 2011.

[37] K. Ognyanova, "Static and dynamic network visualization with R," 2017.

[38] Pew Research Center, "Mobile Fact Sheet," 2018.

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

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

[41] J. T. Kemper, “On the identification of superspreaders for infectious disease,” *Mathematical Biosciences*, vol. 48, no. 1-2, pp. 111–127, 1980.

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

[43] 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.

[44] 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.

[45] 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.

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

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

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

[49] 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.

[50] M. Salathe, M. Kazandjieva, J. W. Lee, P. Levis, M. W. Feldman, and J. H. Jones, “A high-resolution human contact network for infectious disease transmission,” *Proceedings of the National Academy of Sciences*, vol. 107, pp. 22020–22025, dec 2010.

[51] A. Stopczynski, V. Sekara, P. Sapiezynski, A. Cuttone, M. M. Madsen, J. E. Larsen, and S. Lehmann, “Measuring large-scale social networks with high resolution,” *PLoS ONE*, vol. 9, no. 4, 2014.

[52] 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.

[53] 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.

[54] J. Wallinga, P. Teunis, and M. Kretzschmar, “Using data on social contacts to estimate age-specific transmission parameters for respiratory-spread infectious agents,” *American Journal of Epidemiology*, vol. 164, no. 10, pp. 936–944, 2006.

[55] T. A. Reichert, N. Sugaya, D. S. Fedson, W. P. Glezen, L. Simonsen, and M. Tashiro, “The Japanese experience with vaccinating schoolchildren against influenza,” *N Engl J Med*, vol. 344, no. 12, pp. 889–896, 2001.

[56] S. Cauchemez, A.-J. Valleron, P.-Y. Boëlle, A. Flahault, and N. M. Ferguson, “Estimating the impact of school closure on influenza transmission from Sentinel data,” *Nature*, vol. 452, pp. 750–754, apr 2008.

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

[58] 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.

[59] 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.

[60] 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.

[61] 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.

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

[63] 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.

[64] 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.

[65] 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.

[66] 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.

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

[68] 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.

[69] 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.