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

Stephen M. Kissler, Petra Klepac, Maria Tang, Andrew J.K. Conlan and Julia R. Gog

April 7, 2019

Study site

Haslemere is a town in the south of England, in the borough of Waverley, Surrey, with a reported population size of 11,235 people, according to the 2011 UK Census [1]. Crucially for commuters, the town lies on the South Western Railway train line from London Waterloo Station, and sits just off the A3, the major road between London and Portsmouth. Set in the Surrey Hills Area of Outstanding Natural Beauty, the town is popular among walkers and tourists, and has a well-travelled high street [2, 3]. Ultimately, Haslemere was chosen as the study site and epicentre of the national outbreak due to its connectivity and size: its transit links with London make it a plausible site of outbreak establishment, and its size ensures that the town has typical infrastructure features (a rail station, shops, and schools), while being small enough to make recruitment of a significant percentage of the population plausible.

Ethics

Study ethics were approved by an internal review board at the London School of Hygiene and Tropical Medicine.

Recruitment of participants

In the month prior to the data collection period, posters advertising the BBC Pandemic app were disseminated around the town of Haslemere. Representatives from the documentary production company (360 Production) also visited the town on a few occasions to garner support from residents in person.

Data collection

Study participants downloaded the *BBC Pandemic* app onto their smartphone and elected to take part in the 'Haslemere' study. Each user produced a stream of GPS coordinates of up to 1 metre accuracy (though this varies according to individual smartphone GPS tracking capabilities and network connectivity), with recordings taken as frequently as one per five seconds, for three consecutive days. The location recordings were provided by the phone's operating system, which may gather input from satellite-mediated GPS, cell tower triangulation, and the phone's wifi connection to identify the phone's location as accurately as possible. Unlike the 'National' study [4], in which users can record their hourly location data for a 24-hour span of their choice, all Haslemere users participated in the study during the same three days. This allows us to identify actual interactions between app users, rather than just general mobility patterns. Some users dropped into and out of the study at different times during the collection

period, however, and the frequency and accuracy of data collection depended on the strength of the network connection and on whether the phone was moving; a stationary phone would not usually log location updates. Use of the app was restricted to people of at least 16 years of age, or at least 13 years of age with parental consent. A total of 1,272 users logged at least one data point. Since ethical constraints barred us from collecting data on the participants' demographic characteristics, we cannot comment on the extent to which these participants constitute a representative sample of the Haslemere population.

Data cleaning

Since we wished to consider an outbreak in the town of Haslemere, we restrict the raw data to only those users who who spent a significant amount of time within Haslemere. We consider users who logged a location in each of 6 separate hours within the geographic box bound by the coordinates (51.0132, -0.7731) on the south-west and (51.1195, -0.6432) on the north-east, which encompasses the GU27 postcode [5]. This leaves 469 users. To make it possible to directly compare users' locations without having to interpolate between time points, we aggregate each user's data into five-minute bins, spanning from 00:00:00 BST on the first day of the study to 11:59:59 BST on the last day of the study, giving 864 $(60 \div 5 \times 24 \times 3)$ bins. Sometimes, a user does not have a logged time point in a given bin, e.g. if their phone was stationary for a prolonged period of time. For any spans with missing data, we fill the bins with the most recently-recorded location. For missing spans at the beginning of the study, we fill with the user's 'home' location (defined below) if it exists; otherwise, we fill with the first recorded location. We identify a 'home' location for all users with at least 10 logged data points between 22:00 and 07:55 BST on any of the dates, which we define as 'night'. We identify the set of all night-time coordinates logged by each user, and then define the user's home to be 'most local' of those, i.e. the point that has highest number of other night-time points within 20 metres of it. In the case of ties, the earliest 'most local' night-time point is chosen. 'Home' remains undefined for users with fewer than 10 logged night-time points. To ensure users' privacy and to minimise the possibility that individual users could be identified from the outbreak visualisations presented in the BBC Four documentary, we finally filter all users' data to 16 daytime hours only, between 07:00:00 and 22:55:00 BST, since users are likely more identifiable by their night-time locations. There are now 576 $(60 \div 5 \times 16 \times 3)$ time points for each user.

The pairwise distances between users in each 5-minute bin are calculated using the Haversine formula for great-circle geographic distance. A data file containing the pairwise distances at each time point between all users within 50 metres of one another is given in Data S1. This is sufficient to reproduce all results presented in the main text. The location logs themselves cannot be released without potentially compromising the users' privacy.

The Haslemere epidemic simulation model

We model infection using an individual-based susceptible-exposed-infectious (SEI) process. All 468 app users other than the virtual index case begin the simulation as susceptible (S). At each five-minute time step, a susceptible individual may become exposed/infected (E) according to some probability that depends on her/his distance from other infected individuals. Infected individuals become infectious (I) after a fixed (deterministic) amount of time. For the Haslemere outbreak simulation, we assume that individuals remain infectious for the duration of the simulation; there is no recovery within the three days (this is relaxed in the SEIR model, described below).

We specify the probability that a susceptible individual i becomes infected at a given time t as a function of the total force of infection contributed by all infected individuals in the population. We assume that the force of infection decays with distance from the infected individuals

according to some functional form (a 'kernel'), chosen by the modeller. For simplicity, we choose an exponential kernel with a cutoff, so that the force of infection $\lambda_{i,j}(t)$ from infected individual j to a susceptible individual i at time t is

$$\lambda_{i,j}(t) = \begin{cases} ae^{-d_{i,j}/\rho} & d_{i,j} \leq \xi \\ 0 & d_{i,j} > \xi. \end{cases} \quad (1)$$

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. The kernel is depicted in Fig. S1.

Constraints imposed by the documentary narrative required us to choose somewhat unrealistic paramter values. We required an outbreak that would last just three days (or a total of $3 \times 16 = 48$ hours, omitting nighttimes) and infect a high proportion of the population. The speed of a (catastrophic) outbreak can be summarised by the time required to infect 50% of the population. Fig. S2A depicts this time span as a function of a and ρ for simulated epidemics. We sought parameter values such that about 50% of the population would be infected within 24 hours (1.5 days, when restricting to daytimes). For $a = 0.5$, ρ would need to be nearly 50 metres to achieve such a short outbreak. For $a = 1$, $\rho = 10$ metres tends to yield outbreaks for which 50% of the population is infected by day 1.5. As a increases, it is possible to achieve faster outbreaks using smaller values for ρ. For simplicity, we chose $a = 1$ and $\rho = 10$ (Fig. S2, dashed lines) for the Haslemere outbreak, with a cutoff distance of $\xi = 20$ metres. Note that this unrealistically wide range of infection is intended both to speed the outbreak to within the timespan set by the documentary, and to account (as simply as possible) for a range of possible routes of transmission between two individuals. The maximum range of direct transmission of influenza is likely closer to two metres [6, 7]. The wide kernel helps account for the influenza virus' ability to survive on surfaces for limited amounts of time [7], and also for any discrepancy between measured and actual GPS location of the user, both from measurement inaccuracy and from users' movements within the five-minute bin. The underlying model could be refined in many ways to account more subtly for these different effects; but the intention here was to build a minimal model for simulating an outbreak based on the Haslemere dataset. The kernel that corresponds to the parameter values $a = 1$, $\rho = 10$, and $\xi = 20$, the values used to produce the Haslemere outbreak simulation featured in the documentary, is depicted in Fig. S1.

To calculate the probability that susceptible individual i becomes infected at time step t, we first calculate the total force of infection on individual i, $\lambda_i(t) = \sum_j \lambda_{i,j}(t)$. If the total force of infection $\lambda_i(t)$ is interpreted as a survival-analytic hazard, the probability that individual i becomes infected at time t is

$$P_i(t) = 1 - e^{-\lambda_i(t)} \quad (2)$$

as in [8, 9]. We assume that it takes five full time steps (25 minutes) after the time of infection (inclusive) for an individual to become infectious to others (to move from the "E" to the "I" state, in the SEI model).

The extended transmission model

We also consider a more general transmission scenario, in which recovery is possible (an SEIR model). This model is constructed in the same way as the SEI model described above, except the index case is chosen uniform-randomly from the full set of 469 volunteers, and recovery occurs exactly three days (572 time steps) after infection. Upon recovery, an individual can

no longer transmit disease and cannot become infected again. Also, rather than halting the outbreak after three days, we allow the outbreak to continue by looping the data as many times as necessary until no further infected individuals remain.

Individual reproduction number

To calculate the individual reproduction number v_j for user j, we first calculate the probability that individual j infects individual i at some point in the outbreak:

$$P_{i,j} = 1 - \mathsf{Exp}\Big(-\sum_t \lambda_{i,j}(t)\Big), \tag{3}$$

where $\lambda_{i,j}(t)$ is the force of infection on susceptible individual i from infectious individual j at time t, specified by Eq 1. Ignoring all secondary infections, the expected number of infections caused by an infected individual j in an otherwise susceptible population is

$$v_j = \sum_{i \neq j} P_{i,j}. \tag{4}$$

The quantity v_j is the individual reproduction number for person j. The values of v_j range from 0 to over 31. The distribution of v_j for all users is depicted in Fig 3D, separated into the 90^{th}-percentile superspreaders of the Haslemere epidemic (red) and all others (grey). The mean individual reproduction number, $\bar{v}$, is an estimate of the basic reproduction number R_0 [10]. Fig. S2B depicts how $\bar{v} = R_0$ varies as a function of the model parameters a and ρ. For the parameter values used in the featured Haslemere epidemic ($a = 1$, $\rho = 10$), $R_0 = 7.3$, which is higher than most estimates for pandemic influenza [11], but lower than estimates for some other diseases, e.g. measles [12].

Movie S1 depicts the pairwise probability of infection $P_{i,j}$ for a subset of users as a network, where the sum in Eq. 3 is taken over all three days, one day, one hour, and 15 minutes. This illustrates how the time scale on which the infection process is modelled may affect the resulting disease dynamics, and demonstrates the difficulty of adequately capturing the spatiotemporal population structure represented in the Haslemere dataset with one or even a series of static networks.

Empirical estimates of the basic reproduction number

The initial growth rate formula for estimating the basic reproduction number R_0 of an outbreak is

$$R_0 = \frac{r(T_G - T_E)}{\sinh r(T_G - T_E)} e^{rT_G} \tag{5}$$

where T_E is the latent period of the infection (the amount of time it takes a person to transition from 'exposed' to 'infectious'), T_G is the generation interval (the expected time from the onset of one infection to the onset of a secondary infection caused by the first), and r is the exponential growth rate of the cumulative incidence at the start of the outbreak [13]. This approximation assumes fixed latent and infectious periods, and a well-mixed population. To estimate R_0 for outbreaks simulated under the same conditions as the Haslemere epidemic, we set the latent period T_E at 5 time steps or 25 minutes, and the generation interval as the median timespan between each infection and its 'parent', with parent infections assigned randomly with probability weighted by the relative force of infection contributed at the time of infection. The distribution of these generation intervals from the 1,000 simulations used to produce Fig. 5 (main text) is depicted in Fig. S3; its mean is 85.2 time steps (426 minutes) and the middle 90% of observations lie between 71 and 106 time steps (355 and 530 minutes). The exponential

growth rate r is calculated as the slope of the least-squares linear fit to the logged cumulative incidence over time for the first 100 cases of the simulated outbreak, after discarding the first four cases to avoid artefacts from outbreaks that are slow to take off. For outbreaks with fewer than 104 total cases, we do not calculate R_0 using this method.

The final size method for estimating R_0 is

$$R_0 = -\frac{\log(1-f)}{f}. \tag{6}$$

where f is the total fraction of the population that has been infected by the end of the outbreak [14]. This estimate is derived from the standard ordinary differential equation SIR model of disease transmission [15], and therefore also assumes a well-mixed population.

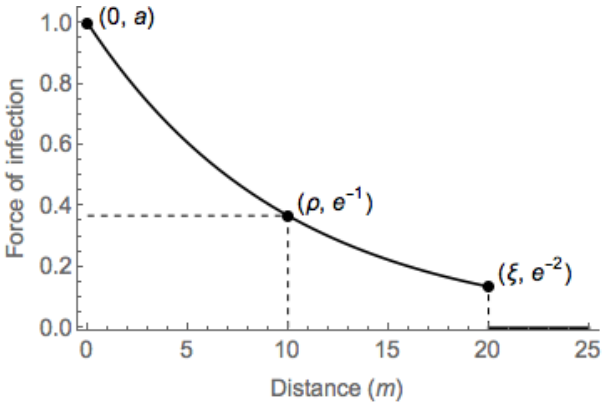

Figure S1: The force of infection λ decays with distance from an infected individual according to the kernel given in Eq 1; an exponential function that intersects the vertical axis at a height of a and decays at rate ρ until a cutoff distance of ξ, after which the force of infection is zero. Parameter values: $a = 1$, $\rho = 10$, and $\xi = 20$. The force of infection halves at a distance of about 7 metres.

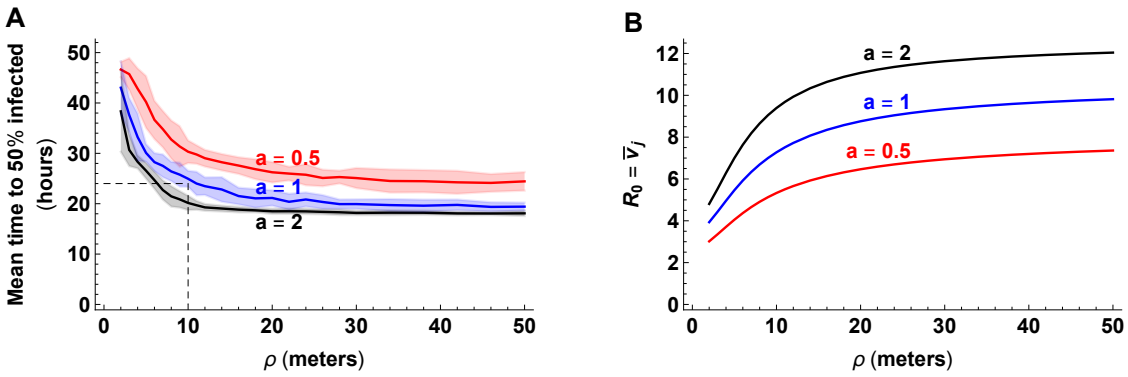

Figure S2: A: Mean time-to-50%-infected for simulated outbreaks as a function of kernel parameters a and ρ. Shaded bands represent the 90% prediction intervals. The horizontal dashed line corresponds to 24 hours, or half of the study period (each day consists of 16 hours, since nighttimes are excluded). For $a = 1$, this corresponds to a characteristic distance ρ of about 10m. B: The basic reproduction number R_0, calculated as the mean individual reproduction number v_j (Eq. 4), as a function of kernel parameters a and ρ.

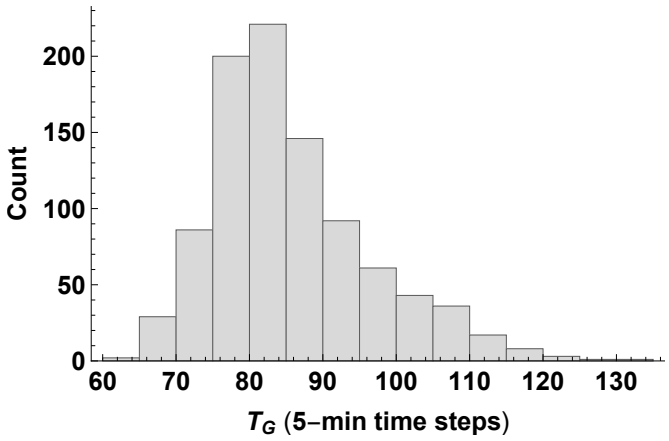

Figure S3: Histogram of median generation intervals from 1,000 simulated epidemics using the extended SEIR transmission model with $a = 1$, $\rho = 10$m, and $\xi = 20$m.

1 Additional files

Movie S1: Pairwise probabilities of infection $P_{i,j}$ (Eq. 3) between the 45 individuals with highest individual reproduction number v_j, where the sum in Eq. 3 is taken over all three days (top left), one day (top right), one hour (bottom left), and 15 minutes (bottom right). Nodes represent individuals, and lines represent probability of infection, such that the opacity of the line is proportional to the probability that an infection would occur in that time interval if one of the individuals were 'infected' and the other were 'susceptible'. The bar along the bottom shows the time span covered by the one-day (top bar), one-hour (middle bar), and 15-minute (bottom bar) networks.

DataS1.csv The "Haslemere dataset", consisting of pairwise distances between users of the BBC Pandemic Haslemere app over time. Each row consists of an encounter (within 50m) between two users. Column 1 gives the time step as an integer value (see DataS2.csv for conversion to real time). Columns 2 and 3 give the user ID numbers. Column 4 gives the distance between the users at that time step, rounded to the nearest metre. Details on the derivation of this dataset are given in the Supplemental Materials and Methods.

DataS2.csv Conversion between the time indices in column 3 of the Haslemere dataset (see Supplemental Materials and Methods) and real time, in British Standard Time (BST).

References

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

[2] Visit Haslemere, "Visit Haslemere," 2018.

[3] Visit Surrey, "Visit Surrey: Haslemere," 2018.

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

[5] Ordnance Survey, "Ordnance Survey: Learn More about Maps and Geography," 2018.

[6] Centers for Disease Control and Prevention, "How Flu Spreads," 2014.

[7] P. R. S. Lagacé-Wiens, E. Rubinstein, and A. Gumel, "Influenza epidemiologypast, present, and future," *Critical Care Medicine*, vol. 38, no. 4, pp. e1–e9, 2010.

[8] 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.

[9] 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.

[10] 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.

[11] P.-Y. Boëlle, S. Ansart, A. Cori, and A.-J. Valleron, "Transmission parameters of the A/H1N1 (2009) influenza virus pandemic: a review," *Influenza and Other Respiratory Viruses*, vol. 5, pp. 306–316, sep 2011.

[12] F. M. Guerra, S. Bolotin, G. Lim, J. Heffernan, S. L. Deeks, Y. Li, and N. S. Crowcroft, "The basic reproduction number (R 0) of measles: a systematic review," *The Lancet Infectious Diseases*, vol. 17, pp. e420–e428, dec 2017.

[13] 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.

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

[15] M. J. Keeling and P. Rohani, *Modeling Infectious Diseases in Humans and Animals*. Princeton University Press, 2011.