medRxiv preprint doi: https://doi.org/10.1101/2020.02.10.20021725; this version posted April 16, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

Beyond R_0: Heterogeneity in secondary infections and probabilistic epidemic forecasting

Laurent Hébert-Dufresne,[1,2,3] Benjamin M. Althouse,[4,5,6] Samuel V. Scarpino,[7,8,9,10,11] and Antoine Allard[3,12]

[1]*Vermont Complex Systems Center, University of Vermont, Burlington, VT 05405*
[2]*Department of Computer Science, University of Vermont, Burlington, VT 05405*
[3]*Département de physique, de génie physique et d'optique, Université Laval, Québec (Québec), Canada G1V 0A6*
[4]*Institute for Disease Modeling, Bellevue, WA*
[5]*University of Washington, Seattle, WA*
[6]*New Mexico State University, Las Cruces, NM*
[7]*Network Science Institute, Northeastern University, Boston, MA, 02115*
[8]*Department of Marine & Environmental Sciences, Northeastern University, Boston, MA, 02115*
[9]*Department of Physics, Northeastern University, Boston, MA, 02115*
[10]*Department of Health Sciences, Northeastern University, Boston, MA, 02115*
[11]*ISI Foundation, Turin, 10126, Italy*
[12]*Centre interdisciplinaire en modélisation mathématique, Université Laval, Québec (Québec), Canada G1V 0A6*

The basic reproductive number — R_0 — is one of the most common and most commonly misapplied numbers in public health. Although often used to compare outbreaks and forecast pandemic risk, this single number belies the complexity that two different pathogens can exhibit, even when they have the same R_0 [1–3]. Here, we show how to predict outbreak size using estimates of the distribution of secondary infections, leveraging both its average R_0 and the underlying heterogeneity. To do so, we reformulate and extend a classic result from random network theory [4] that relies on contact tracing data to simultaneously determine the first moment (R_0) and the higher moments (representing the heterogeneity) in the distribution of secondary infections. Further, we show the different ways in which this framework can be implemented in the data-scarce reality of emerging pathogens. Lastly, we demonstrate that without data on the heterogeneity in secondary infections for emerging infectious diseases like COVID-19, the uncertainty in outbreak size ranges dramatically. Taken together, our work highlights the critical need for contact tracing during emerging infectious disease outbreaks and the need to look beyond R_0 when predicting epidemic size.

I. INTRODUCTION

In 1918, a typical individual infected with influenza transmitted the virus to between one and two of their social contacts [5], giving a value of the basic reproductive number – R_0, the number of secondary infections in a completely susceptible population – between one and two. These are similar to values of R_0 for the 2014 West Africa Ebola virus outbreak, but most individuals infected with Ebola virus gave rise to zero additional infections, while a few gave rise to more than 10 [6, 7]. Moreover, Ebola virus disease infected a tenth of one percent of the number of individuals believed to have been infected by the 1918 Influenza virus [8, 9]. While improvements in healthcare and public health measures, as well as changes in human behavior, partially explain the massive discrepancy between Ebola virus disease in 2014 and influenza in 1918 [10], there is another critical difference between these two diseases: heterogeneity in the number secondary cases resulting from a single infected individual. Here, we demonstrate analytically that quantifying the variability in the number of secondary infections is critically important for quantifying the transmission risk of novel pathogens.

The basic reproduction number of an epidemic, R_0, is the expected number of secondary cases (note, we use the word "case" in a generic sense to represent any infection, even if too mild to meet the clinical case definition) produced by a primary case over the course of their infectious period in a completely susceptible population [11]. It is a simple metric that is commonly used to describe and compare the transmissibilty of emerging and endemic pathogens [12]. If $R_0 = 2$, one case turns to two, on average, and two turn to four as the epidemic grows. Conversely, the epidemic will die out if $R_0 < 1$.

Almost 100 years ago, work from Kermack and McKendrick [13–15] first demonstrated how to estimate the final size of an epidemic. Specifically, they considered a scenario such that: *(i)* the disease results in complete immunity or death, *(ii)* all individuals are equally susceptible, *(iii)* the disease is transmitted in a closed population, *(iv)* contacts occur according to the law of mass-action, *(v)* and the population is large enough to justify a deterministic analysis. Under these assumptions, Kermack and McKendrick show that an epidemic with a given R_0 will infect a fixed fraction $R(\infty)$ of the susceptible population by solving

$$R(\infty) = -\frac{1}{R_0} \ln\left[1 - R(\infty)\right] \ . \tag{1}$$

This solution describes a final outbreak size equal to 0 when $R_0 \leq 1$ and increasing roughly as $1 - \exp(-R_0)$ when $R_0 > 1$. Therefore, a larger R_0 leads to a larger outbreak which infects the entire population in the limit $R_0 \to \infty$. This direct relationship between R_0 and the final epidemic size is at the core of the conventional wisdom that a larger R_0 will cause a larger outbreak.

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

medRxiv preprint doi: https://doi.org/10.1101/2020.02.10.20021725; this version posted April 16, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

Unfortunately, the equation relating R_0 to final outbreak size from Kermack and McKendrick is only valid when all the above assumptions hold, which is rarely the case in practice.

As a result, relying on R_0 alone is often misleading when comparing different pathogens or outbreaks of the same pathogen in different settings [1–3]. This is especially critical considering that many outbreaks are not shaped by the "average" individuals but rather by a minority of super-spreading events [1, 16]. To more fully quantify how heterogeneity in the number of secondary infections affects outbreak size, we turn towards network epidemiology and derive an equation for the total number of infected individuals using all moments of the distribution of secondary infections.

II. RANDOM NETWORK ANALYSIS

Random network theory allows us to relax some of assumptions made by Kermack and McKendrick, mainly to account for heterogeneity and stochasticity in the number of secondary infections caused by a given individual. We first follow the analysis of Ref. [4] and define

$$G_0(x) = \sum_{k=0}^{\infty} p_k x^k \tag{2}$$

as the probability generating function (PGF) of the distribution of the number on contacts $\{p_k\}$ individuals have (the *degree* distribution). When following a random contact (an *edge*), we define the *excess* degree as the number of other edges around that node reached via one of its edges. Because an edge is k times more likely to reach a node of degree k than a node of degree 1, the excess degree distribution is generated by

$$G_1(x) = \frac{G_0'(x)}{G_0'(1)} = \frac{1}{\langle k \rangle} \sum_{k=1}^{\infty} k p_k x^{k-1} \tag{3}$$

where $\langle k \rangle$ is the average degree and acts as a normalisation constant, and $G_0'(x)$ denotes the derivative of $G_0(x)$ with respect to x.

We now assume that the network in question is the network of all edges that *would* transmit a disease if given the chance. Consequently, $G_1(x)$ generates the number of secondary infections that individual nodes would cause if infected. And, if we infect a random node as the patient zero, its entire connected component (a maximal subset of nodes between which paths exists between all pairs of nodes) will be infected. To calculate the largest possible epidemic, we thus look for the size of the giant connected component (GCC).

To calculate the size of the GCC, we first look for the probability u that following a random edge leads to a node *not* part of the GCC. For that node to not be part of the GCC, all of its excess edges must also not lead to the GCC. This simple observation leads to the self-consistent equation

$$u = \frac{1}{\langle k \rangle} \sum_{k=1}^{\infty} k p_k u^{k-1} = G_1(u) \; . \tag{4}$$

The size of the GCC is a fraction of the full population N that we will denote $R(\infty)$ because it corresponds to the potential, macroscopic, outbreak size. Noting that a node of degree k is *not* in the GCC with probability u^k, $R(\infty)$ can be calculated by counting all nodes except those with no edges leading to the GCC,

$$R(\infty) = 1 - G_0(u) \; . \tag{5}$$

When dealing with data on the distribution of secondary infections, $G_0(x)$ has one degree of freedom remaining, p_0, which we set by assuring that the number of infections caused by patient zero is smaller or equal to R_0 but not greater (see Methods). The resulting solution for $R(\infty)$ is exact in the limit of infinite population size.

III. RESULTS

The network approach naturally accounts for heterogeneity, meaning that some individuals will cause more infections than others. The network approach also accounts for stochasticity explicitly: Even with $R_0 > 1$, there is a probability $1 - R(\infty)$ that patient zero lies outside of the giant outbreak and therefore only leads to a small outbreak that does not invade the population. However, the analysis in terms of PGFs is obviously more involved than simply assuming mass-action mixing and solving Eq. (1). In fact, the PGF $G_0(x)$ requires a full distribution of secondary cases per primary case, which will in practice involve the specification of a high-order polynomial.

medRxiv preprint doi: https://doi.org/10.1101/2020.02.10.20021725; this version posted April 16, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

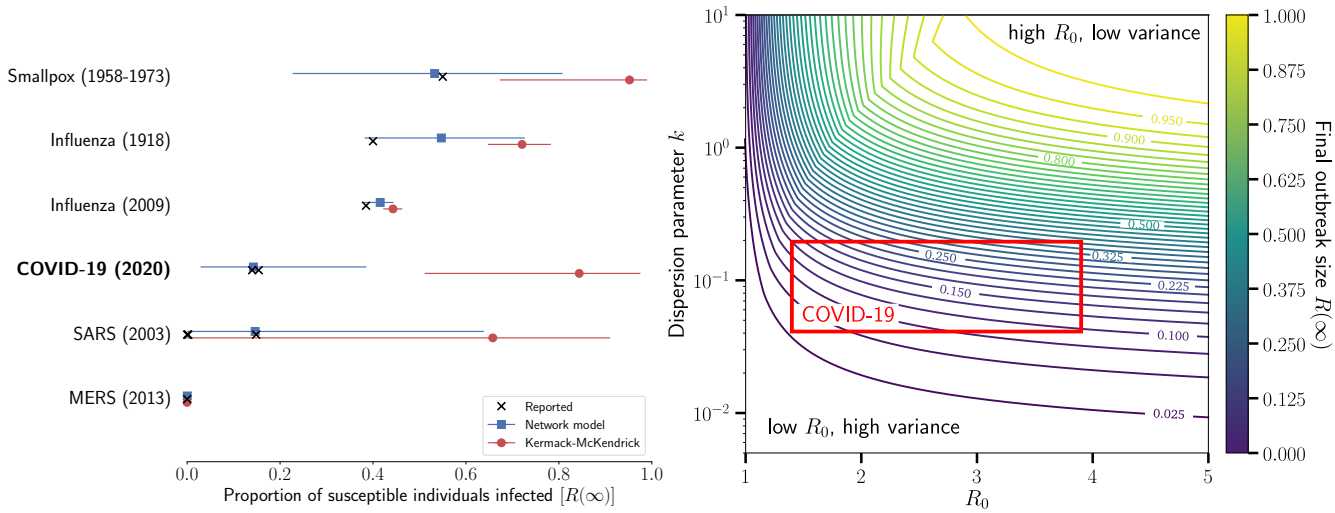

FIG. 1. (left) Using published estimates of R_0 and the dispersion parameter, k, we estimated the total outbreak size for six different diseases. The confidence intervals span the range of uncertainty reported for R_0 and k. The black markers show reported total outbreak sizes (total proportion of susceptible individuals infected) for each disease. For influenza, we report the estimated proportion of school-aged children infected. The blue squares show the estimated proportion infected obtained with Eq. (5). The red circles are the estimated proportion infected using the method developed by Kermack and McKendrick, i.e., Eq. (1). (right) Final size of outbreaks with different R_0 and distributions of secondary cases. We use a negative binomial distribution of secondary cases and scan a realistic range of parameters. The range of parameters corresponding to current estimates for COVID-19 is highlighted by a red box. Most importantly, with fixed average, the dispersion parameter is inversely proportional to the variance of the underlying distribution of second cases. See Table I and `https://github.com/Emergent-Epidemics/beyond_R0` for additional information and to access the data.

To clarify and potentially simplify the approach, we propose to reformulate the classic network model in terms of the cumulant generating function (CGF) of secondary cases. The CGF $K(y)$ of a random variable X can be written as $K(y) = \sum \kappa_n y^n/n!$ where κ_n are the cumulants of the distribution of secondary infections. These are useful because the cumulants are easier to interpret, i.e., κ_1 is simply the average number of secondary cases R_0, κ_2 is the variance, κ_3 is related to the skewness and κ_4 to the kurtosis of the full distribution, etc. By definition, a PGF $G(x)$ of a random variable is linked to $K(y)$ through $G(x) = \exp[K(\ln x)]$. Therefore, we can replace the PGF $G_1(x)$ for the distribution of secondary infections by a function in terms of the cumulants of that distribution.

A. Analysis of cumulants and derivation of Kermack-McKendrick

We can easily derive Kermack and McKendrick's result from this framework since their solution assumes a well-mixed population, which corresponds to a Poisson distribution of secondary infections. We first re-write $G_1(x)$ in terms of the cumulants κ_n as

$$G_1(x) = \exp\left[\sum_{n=1}^{\infty} \frac{1}{n!}\kappa_n (\ln x)^n\right] , \tag{6}$$

which is a particular convenient representation for a Poisson distribution because its cumulants $\kappa_n = R_0$ for all $n > 0$. Moreover, since $G_0(x) = G_1(x)$ in the Poisson case, the final outbreak size of the Kermack-McKendrick analysis will be set by $u_{\rm KM} = G_1(u_{\rm KM})$, or

$$\begin{aligned} u_{\rm KM} &= \exp\left[\sum_{n=1}^{\infty} \frac{1}{n!} R_0 (\ln u_{\rm KM})^n\right] = \exp\Big[R_0 (u_{\rm KM} - 1)\Big] ; \\ \hookrightarrow \quad & R_{\rm KM}(\infty) = 1 - \exp[R_0 (u_{\rm KM} - 1)] = 1 - \exp[-R_0 R_{\rm KM}(\infty)] \end{aligned} \tag{7}$$

Taking the logarithm of the exponential term from this last equation yields Kermack and McKendrick's formula.

The solution to $u = G_1(u)$ gives the probability that every infection caused by patient zero fails to generate an epidemic. For

medRxiv preprint doi: https://doi.org/10.1101/2020.02.10.20021725; this version posted April 16, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

Disease	Location	Year	Prop. Infect.	R_0	k	Ref.
MERS	Global	2013	0.00	0.47 (0.29–0.80)§	0.26 (0.09–1.24)§	[17, 18]
SARS	Global	2003	0.00–0.15	1.63 (0.54–2.65)*	0.16 (0.11–0.64)*	[1, 19, 20]
Smallpox	Europe	1958-1973	0.55	3.19 (1.66–4.62)*	0.37 (0.26–0.69)*	[1, 21]
Influenza	Baltimore (USA)	1918	0.40	1.77 (1.61–1.95)§	0.94 (0.59–1.72)§	[22, 23]
Influenza	Italy	2009	0.39	1.321 (1.299–1.343)§	8.092 (5.170–11.794)§	[24, 25]
COVID-19	Global	2020	0.14, 0.154	2.5 (1.4–12)§	0.1 (0.04–0.2)§	[26–29]

TABLE I. Estimates for R_0 and for the negative binomial distribution dispersion parameter, k, used on Fig. 1 (§ and $\star$ respectively denote 95% and 90% confidence intervals). Proportion of susceptible individuals infected as reported in either the literature or by the US CDC. For SARS, the proportion of infected was taken from serosurveys among wild animal handlers (0.15) and among health-care workers (<0.01) [19]. For influenza (2009), we took data on school-aged children. For COVID-19, the proportion of infected was taken from a serosurvey in the municipality of Gangelt, Germany [28] and from universal testing in all obstetrical patients presenting for delivery at two hospitals [29]. Note that the estimates the proportion of infected individuals, for R_0 and for k were not necessarily inferred from the same populations. Such information is rarely, if ever, available for a same outbreak, unfortunately.

more general distributions, it is useful to rewrite Eq. (6) as

$$\begin{aligned} u = G_1(u) &= \exp\left[\sum_{n=1}^{\infty} \frac{1}{n!}\kappa_n (\ln u)^n\right] \\ &= \exp\left[R_0|\ln u| - \frac{1}{2}\sigma^2|\ln u|^2 + \frac{1}{6}\kappa_3|\ln u|^3 - \frac{1}{24}\kappa_4|\ln u|^4 \ldots\right] \end{aligned} \tag{8}$$

to highlight its alternating nature because $\ln u$ is negative (u is a probability) such that its n-th power is positive when n is even and negative when n is odd.

The alternating sign of contribution from high-order moments in Eq. (8) can be interpreted as follows. A disease needs a high average number of secondary infections (high $\kappa_1 = R_0$) to spread, but given that average, a disease with small variance in secondary infections will spread much more reliably and be less likely to stochastically die out. Given a variance, a disease with high skewness (i.e., with positive deviation contributing to most of the variance) will be more stable than a disease with negative skewness (i.e. with most deviations being towards small secondary infections). Given a skewness, a disease will be more stable if it has frequent small positive deviations rather than infrequent large deviations — hence a smaller kurtosis — as stochastic die out could easily occur before any of those large infrequent deviations occur.

Our re-interpretation already highlights a striking result: Higher moments of the distribution of secondary cases can lead a disease with a lower R_0 to invade more easily a population and to reach a larger final outbreak size than a disease with a higher R_0. Taking into account the contribution of these higher moments also yields better estimates for the final size of outbreaks, as we now show.

B. Comparison of estimators to empirical data

We now compare the final outbreak size estimates from Eq. (1) (Kermack and McKendrick) to estimates from Eq. (5) (network model) with a negative binomial offspring distribution (see Methods and Table I). As predicted, Fig. 1(left) shows that the Kermack and McKendrick formulation consistently and significantly over-predicts the outbreak size across six different pathogens where we could find confidence interval estimates for R_0 and for the negative binomial over-dispersion parameter (k). Our approach produces estimates of the total outbreak size, which are consistent with outbreaks where no vaccine was available (smallpox in unvaccinated populations, the 1918 influenza pandemic, and school children prior to the availability of the 2009 H1N1 vaccine). Clearly, once interventions are put in place and/or substantial behavioral change occurs, all methods that do not account for these effects will over-estimate the total outbreak size [30]. Nevertheless, our approach provides a much more reasoned estimate of the total risk to any given population, and predictions very close to the most recent seropositivity estimates for the COVID-19 outbreak in a German Municipality [28] and in obstetrical patients presenting for delivery [29], as well as for SARS among wild animal handlers (other smaller estimates correspond to health-care workers) [19].

IV. DISCUSSION

From re-emerging pathogens like yellow fever and measles to emerging threats like Middle East Respiratory Syndrome coronavirus and Ebola, the World Health Organization monitored 119 different infectious disease outbreaks in 2019 alone [31]. For

medRxiv preprint doi: https://doi.org/10.1101/2020.02.10.20021725; this version posted April 16, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

each of these outbreaks, predicting both the epidemic potential and the most likely number of cases is critically important for efficient and effective responses. This need for rapid situational awareness is why R_0 is so widely used in public health. However, our main analysis shows that not only is R_0 insufficient in fully determining the final size of an outbreak, but having a larger outbreak with a lower R_0 is relatively easy considering the randomness associated with most transmission events and the heterogeneity of physical contacts. To address the need for rapid quantification of risk, while acknowledging the shortcomings of R_0, we use network science methods to derive both the probability of an epidemic and its final size.

These results are not without important caveats. Specifically, we must remember that distributions of secondary cases, just like R_0 itself, are just as much a product of a pathogen as of the population in which it spreads. For example, aspects of the social contact network [32], metapopulation structure [33], human mobility [34], adaptive behavior [35], and even other pathogens [36, 37], all interact to cause complex patterns of disease emergence, spread, and persistence. Therefore, great care must be taken when using any of these tools to compare outbreaks or to inform current events with past data.

Figure 1(left) only used a handful of known outbreaks to validate the different approaches because data on secondary cases are rare. In practice, three types of data could potentially be used in real time to improve predictions by considering secondary case heterogeneity. First, contact tracing data whose objective is to identify people who may have come into contact with an infectious individual. While mostly a preventive measure to identify cases before complications, it directly informs us about potential secondary cases caused by a single individual, and therefore provides us with an estimate for $G_1(x)$. Both for generating accurate predictions of epidemic risk and controlling the outbreak, it is vital to begin contact tracing before numerous transmission chains become widely distributed across space [38, 39].

Second, viral genome sequences provide information on both the timing of the outbreak [40] and structure of secondary cases [41]. For example, methods exist to reconstruct transmission trees for sampled sequences using simple mutational models to construct a likelihood for a specific transmission tree [42, 43] and translate coalescent rates into key epidemiological parameters [44, 45]. Despite the potential for genome sequencing to revolutionize outbreak response, the global public health community still struggles to coordinate data sharing across international borders, between academic researchers, and with private companies [46–48].

Third, early incidence data can be leveraged to infer parameters of the secondary case distribution through comparison with simulations. Comparing the output of agent-based simulations with reported incidence can be used to effectively sample a joint posterior distribution over R_0 and dispersion parameter k. This approach was used by most studies referenced in Table I. Most importantly, these simulations need not be run over long periods of time to predict final outbreak size. Instead, they only need to be run over enough early data to infer the parameter estimates that are then fed into our network model to compute the final outbreak size.

As for COVID-19, Fig. 1(right) shows how the width of the confidence interval on our prediction for the final outbreak size mostly stems from uncertainty in the heterogeneity of secondary infections; i.e., the dispersion parameter k. With limited heterogeneity, our predictions would have been closer to classic mass-action forecasts and the current pandemic of COVID-19 would likely have been a consequence of not only R_0, but of the homogeneity of secondary infections: each new cases steadily leading to additional infections. Thankfully, with recent large estimates for its heterogeneity, the observed transmission could be mostly maintained by so-called "super-spreading events", which could be easier to manage with contact tracing, screening and infection control [49].

In conclusion, we reiterate that when accounting for the full distribution of secondary cases caused by an infected individual, there is no direct relationship between R_0 and the size of an outbreak. We also stress that both R_0 and the full secondary case distribution are not properties of the disease itself, but are instead set by properties of the pathogen, the host population and the context of the outbreak. Nevertheless, we provide a straightforward methodology for translating estimates of transmission heterogeneity into epidemic forecasts. Altogether, predicting outbreak size based on early data is an incredibly complex challenge but one that is increasingly within reach due to new mathematical analyses and faster communication of public health data.

ACKNOWLEDGMENTS

L.H.-D. acknowledges support from the National Institutes of Health 1P20 GM125498-01 Centers of Biomedical Research Excellence Award. B.M.A. is supported by Bill and Melinda Gates through the Global Good Fund. S.V.S. is supported by startup funds provided by Northeastern University. A.A. acknowledges financial support from the Sentinelle Nord initiative of the Canada First Research Excellence Fund and from the Natural Sciences and Engineering Research Council of Canada (project 2019-05183).

medRxiv preprint doi: https://doi.org/10.1101/2020.02.10.20021725; this version posted April 16, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

METHODS

The results presented from our network model assume the number of secondary infections to be distributed according to a negative binomial distribution parametrized by its average R_0 and dispersion k [1]. Its probability generating function (PGF) is

$$G_1(x) = \sum_{n=0}^{\infty} \binom{n+k-1}{n} \left[\frac{R_0}{R_0+k}\right]^n \left[1 - \frac{R_0}{R_0+k}\right]^k x^n = \left[1 + \frac{R_0}{k}(1-x)\right]^{-k} . \tag{9}$$

The network theory formalism presented in the main text requires the specification of the PGF $G_0(x)$ whose related to $G_1(x)$ via

$$G_1(x) = \frac{G_0'(x)}{G_0'(1)} \tag{10}$$

where the prime ($\prime$) denotes the first derivative with respect to x. Specifying $G_1(x)$ therefore fixes $G_0(x)$ up to a constant and to a multiplicative factor. Without loss of generality, we set

$$G_0(x) = p_0 + (1-p_0) g_0(x) \tag{11}$$

with $0 \leq p_0 \leq 1$, $g_0(0) = 0$ and $g_1(1) = 1$. Equation (10) becomes

$$G_1(x) = \frac{g_0'(x)}{g_0'(1)} , \tag{12}$$

from which we compute

$$g_0(x) = \int g_0'(x)dx = g_0'(1) \int G_1(x)dx = -\frac{k g_0'(1)}{R_0(1-k)} \left[1 + \frac{R_0}{k}(1-x)\right]^{1-k} + C , \tag{13}$$

with $k \neq 1$, and where C and $g_0'(1)$ are fixed by imposing $g_0(0) = 0$ and $g_1(1) = 1$. Rearranging the terms, we find that

$$g_0(x) = \frac{1 - \left[1 - \frac{R_0 x}{R_0+k}\right]^{1-k}}{1 - \left[\frac{k}{R_0+k}\right]^{1-k}} , \tag{14}$$

from which we finally obtain

$$G_0(x) = p_0 + (1-p_0) \frac{1 - \left[1 - \frac{R_0 x}{R_0+k}\right]^{1-k}}{1 - \left[\frac{k}{R_0+k}\right]^{1-k}} \tag{15}$$

with $k \neq 1$. The case $k = 1$ must be treated separately and yields

$$G_0(x) = p_0 + (1-p_0) \left[1 - \frac{\ln[1 + R_0(1-x)]}{\ln[1+R_0]}\right] . \tag{16}$$

From Eqs. (15) and (16), we find that the average number of secondary infections caused by the patient zero is

$$G_0'(1) = (1-p_0) \frac{(1-k)R_0}{k} \frac{1}{\left[\frac{k}{R_0+k}\right]^{k-1} - 1} \tag{17}$$

if $k \neq 1$, and

$$G_0'(1) = (1-p_0) \frac{R_0}{\ln[1+R_0]} \tag{18}$$

if $k = 1$. The average number of secondary infections caused by patient zero can therefore be greater or smaller than R_0. Since patient zero should not be expected to create *more* secondary cases than the next generation of infections, we set the value of $p_0 \in [0, 1]$ such that $G_0'(1)$ is as close as possible to R_0 whenever $G_0'(1) > R_0$.

medRxiv preprint doi: https://doi.org/10.1101/2020.02.10.20021725; this version posted April 16, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

A large-scale epidemic is predicted by this framework [4] if

$$G_1'(1) = R_0 > 1 \ , \tag{19}$$

as in the analysis by Kermack and McKendrick [13–15]. Its size, $R(\infty)$, is computed with $G_0(x)$ as

$$R(\infty) = 1 - G_0(u) \tag{20}$$

where u is the solution of

$$u = G_1(u) \tag{21}$$

which we solve using the relaxation method [50] with an initial condition randomly chosen in the open interval $(0, 1)$.

[1] 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 **438**, 355 (2005).

[2] S. Bansal, B. T. Grenfell, and L. A. Meyers, When individual behaviour matters: homogeneous and network models in epidemiology, J. R. Soc. Interface **4**, 879 (2007).

[3] Z. Vizi, I. Z. Kiss, J. C. Miller, and G. Röst, A monotonic relationship between the variability of the infectious period and final size in pairwise epidemic modelling, arXiv (2017), arXiv:1712.06026.

[4] M. E. J. Newman, S. H. Strogatz, and D. J. Watts, Random graphs with arbitrary degree distributions and their applications, Phys. Rev. E **64**, 026118 (2001).

[5] M. Biggerstaff, S. Cauchemez, C. Reed, M. Gambhir, and L. Finelli, Estimates of the reproduction number for seasonal, pandemic, and zoonotic influenza: a systematic review of the literature, BMC Infect. Dis. **14**, 480 (2014).

[6] WHO Ebola Response Team, Ebola Virus Disease in West Africa — The First 9 Months of the Epidemic and Forward Projections, N. Engl. J. Med. **371**, 1481 (2014).

[7] C. L. Althaus, Estimating the Reproduction Number of Ebola Virus (EBOV) During the 2014 Outbreak in West Africa, PLOS Curr. 10.1371/currents.outbreaks.91afb5e0f279e7f29e7056095255b288 (2014).

[8] C. E. Mills, J. M. Robins, and M. Lipsitch, Transmissibility of 1918 pandemic influenza, Nature **432**, 904 (2004).

[9] J. Kaner and S. Schaack, Understanding Ebola: the 2014 epidemic, Global. Health **12**, 53 (2016).

[10] M. C. J. Bootsma and N. M. Ferguson, The effect of public health measures on the 1918 influenza pandemic in U.S. cities, Proc. Natl. Acad. Sci. USA **104**, 7588 (2007).

[11] O. Diekmann, J. A. J. Metz, and J. A. P. Heesterbeek, The legacy of Kermack and McKendrick, in *Epidemic Model. Their Struct. Relat. to Data*, edited by D. Mollison (Cambridge University Press, 1995) pp. 95–115.

[12] K. Sheikh, D. Watkins, J. Wu, and M. Gröndahl, How Bad Will the Coronavirus Outbreak Get? Here Are 6 Key Factors, The New York Times (2020).

[13] W. O. Kermack and A. G. McKendrick, A Contribution to the Mathematical Theory of Epidemics, Proc. R. Soc. A **115**, 700 (1927).

[14] W. O. Kermack and A. G. McKendrick, Contributions to the Mathematical Theory of Epidemics. II. The Problem of Endemicity, Proc. R. Soc. A **138**, 55 (1932).

[15] W. O. Kermack and A. G. McKendrick, Contributions to the Mathematical Theory of Epidemics. III. Further Studies of the Problem of Endemicity, Proc. R. Soc. A **141**, 94 (1933).

[16] L. A. Meyers, B. Pourbohloul, M. E. J. Newman, D. M. Skowronski, and R. C. Brunham, Network theory and SARS: Predicting outbreak diversity., J. Theor. Biol. **232**, 71 (2005).

[17] Z. A. Memish, A. Assiri, M. Almasri, R. F. Alhakeem, A. Turkestani, A. A. Al Rabeeah, J. A. Al-Tawfiq, A. Alzahrani, E. Azhar, H. Q. Makhdoom, W. H. Hajomar, A. M. Al-Shangiti, and S. Yezli, Prevalence of MERS-CoV Nasal Carriage and Compliance With the Saudi Health Recommendations Among Pilgrims Attending the 2013 Hajj, J. Infect. Dis. **210**, 1067 (2014).

[18] A. J. Kucharski and C. L. Althaus, The role of superspreading in Middle East respiratory syndrome coronavirus (MERS-CoV) transmission, Eurosurveillance **20**, pii=21167 (2015).

[19] G. M. Leung, W. W. Lim, L.-M. Ho, T.-H. Lam, A. C. Ghani, C. A. Donnelly, C. Fraser, S. Riley, N. M. Ferguson, R. M. Anderson, and A. J. Hedley, Seroprevalence of IgG antibodies to SARS-coronavirus in asymptomatic or subclinical population groups, Epidemiology and Infection **134**, 211 (2006).

[20] S. R. Quah and L. Hin-Peng, Crisis Prevention and Management during SARS Outbreak, Singapore, Emerg. Infect. Dis. **10**, 364 (2004).

[21] T. M. Mack, D. B. Thoma, A. Ali, and M. M. Khan, Epidemiology of smallpox in West Pakistan, Am. J. Epidemiol. **95**, 157 (1972).

[22] J. K. Taubenberger and D. M. Morens, 1918 Influenza: the Mother of All Pandemics, Emerg. Infect. Dis. **12**, 15 (2006).

[23] C. Fraser, D. A. T. Cummings, D. Klinkenberg, D. S. Burke, and N. M. Ferguson, Influenza Transmission in Households During the 1918 Pandemic, Am. J. Epidemiol. **174**, 505 (2011).

[24] H. Kelly, H. A. Peck, K. L. Laurie, P. Wu, H. Nishiura, and B. J. Cowling, The Age-Specific Cumulative Incidence of Infection with Pandemic Influenza H1N1 2009 Was Similar in Various Countries Prior to Vaccination, PLOS ONE **6**, e21828 (2011).

[25] I. Dorigatti, S. Cauchemez, A. Pugliese, and N. M. Ferguson, A new approach to characterising infectious disease transmission dynamics from sentinel surveillance: Application to the Italian 2009–2010 A/H1N1 influenza pandemic, Epidemics **4**, 9 (2012).

medRxiv preprint doi: https://doi.org/10.1101/2020.02.10.20021725; this version posted April 16, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

[26] Q. Li, X. Guan, P. Wu, X. Wang, L. Zhou, Y. Tong, R. Ren, K. S. M. Leung, E. H. Y. Lau, J. Y. Wong, X. Xing, N. Xiang, Y. Wu, C. Li, Q. Chen, D. Li, T. Liu, J. Zhao, M. Liu, W. Tu, C. Chen, L. Jin, R. Yang, Q. Wang, S. Zhou, R. Wang, H. Liu, Y. Luo, Y. Liu, G. Shao, H. Li, Z. Tao, Y. Yang, Z. Deng, B. Liu, Z. Ma, Y. Zhang, G. Shi, T. T. Y. Lam, J. T. Wu, G. F. Gao, B. J. Cowling, B. Yang, G. M. Leung, and Z. Feng, Early Transmission Dynamics in Wuhan, China, of Novel Coronavirus–Infected Pneumonia, N. Engl. J. Med. 10.1056/NEJMoa2001316 (2020).

[27] A. Endo, Centre for the Mathematical Modelling of Infectious Diseases COVID-19 Working Group, S. Abbott, A. J. Kucharski, and S. Funk, Estimating the overdispersion in COVID-19 transmission using outbreak sizes outside {China}, Wellcome Open Res. **5**, 67 (2020).

[28] H. Streeck, G. Hartmann, M. Exner, and M. Schmid, Preliminary result and conclusions of the COVID-19 case cluster study (Gangelt Municipality), online preprint (2020).

[29] D. Sutton, K. Fuchs, M. D'Alton, and D. Goffman, Universal Screening for SARS-CoV-2 in Women Admitted for Delivery, New England Journal of Medicine , NEJMc2009316 (2020).

[30] C. Eksin, K. Paarporn, and J. S. Weitz, Systematic biases in disease forecasting – The role of behavior change, Epidemics **27**, 96 (2019).

[31] WHO disease outbreaks by year: 2019, `https://www.who.int/csr/don/archive/year/2019/en/` (2019), accessed: 2020-02-09.

[32] Y. Moreno, R. Pastor-Satorras, and A. Vespignani, Epidemic outbreaks in complex heterogeneous networks, Eur. Phys. J. B **26**, 521 (2002).

[33] V. Colizza and A. Vespignani, Epidemic modeling in metapopulation systems with heterogeneous coupling pattern: Theory and simulations, J. Theor. Biol. **251**, 450 (2008).

[34] 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, Proc. Natl. Acad. Sci. USA **112**, 11887 (2015).

[35] S. V. Scarpino, A. Allard, and L. Hébert-Dufresne, The effect of a prudent adaptive behaviour on disease transmission, Nat. Phys. **12**, 1042 (2016).

[36] L. Hébert-Dufresne and B. M. Althouse, Complex dynamics of synergistic coinfections on realistically clustered networks, Proc. Natl. Acad. Sci. USA **112**, 10551 (2015).

[37] L. Hébert-Dufresne, S. V. Scarpino, and J.-G. Young, Interacting contagions are indistinguishable from social reinforcement, arXiv (2019), arXiv:1906.01147.

[38] R. S. Dhillon and D. Srikrishna, When is contact tracing not enough to stop an outbreak?, Lancet Infect. Dis. **18**, 1302 (2018).

[39] D. Klinkenberg, C. Fraser, and H. Heesterbeek, The Effectiveness of Contact Tracing in Emerging Epidemics, PLOS ONE **1**, e12 (2006).

[40] G. J. D. Smith, D. Vijaykrishna, J. Bahl, S. J. Lycett, M. Worobey, O. G. Pybus, S. K. Ma, C. L. Cheung, J. Raghwani, S. Bhatt, J. S. M. Peiris, Y. Guan, and A. Rambaut, Origins and evolutionary genomics of the 2009 swine-origin H1N1 influenza A epidemic, Nature **459**, 1122 (2009).

[41] S. V. Scarpino, A. Iamarino, C. Wells, D. Yamin, M. Ndeffo-Mbah, N. S. Wenzel, S. J. Fox, T. Nyenswah, F. L. Altice, A. P. Galvani, L. A. Meyers, and J. P. Townsend, Epidemiological and Viral Genomic Sequence Analysis of the 2014 Ebola Outbreak Reveals Clustered Transmission, Clin. Infect. Dis. **60**, 1079 (2015).

[42] T. Jombart, A. Cori, X. Didelot, S. Cauchemez, C. Fraser, and N. Ferguson, Bayesian Reconstruction of Disease Outbreaks by Combining Epidemiologic and Genomic Data, PLOS Comput. Biol. **10**, e1003457 (2014).

[43] F. Campbell, A. Cori, N. Ferguson, and T. Jombart, Bayesian inference of transmission chains using timing of symptoms, pathogen genomes and contact data, PLOS Comput. Biol. **15**, e1006930 (2019).

[44] E. M. Volz, K. Koelle, and T. Bedford, Viral Phylodynamics, PLOS Comput. Biol. **9**, e1002947 (2013).

[45] R. Bouckaert, T. G. Vaughan, J. Barido-Sottani, S. Duchêne, M. Fourment, A. Gavryushkina, J. Heled, G. Jones, D. Kühnert, N. De Maio, M. Matschiner, F. K. Mendes, N. F. Müller, H. A. Ogilvie, L. du Plessis, A. Popinga, A. Rambaut, D. Rasmussen, I. Siveroni, M. A. Suchard, C.-H. Wu, D. Xie, C. Zhang, T. Stadler, and A. J. Drummond, BEAST 2.5: An advanced software platform for Bayesian evolutionary analysis, PLOS Comput. Biol. **15**, e1006650 (2019).

[46] J. Gardy, N. J. Loman, and A. Rambaut, Real-time digital pathogen surveillance — the time is now, Genome Biol. **16**, 155 (2015).

[47] S. Van Puyvelde and S. Argimon, Sequencing in the time of Ebola, Nat. Rev. Microbiol. **17**, 5 (2019).

[48] N. D. Grubaugh, J. T. Ladner, P. Lemey, O. G. Pybus, A. Rambaut, E. C. Holmes, and K. G. Andersen, Tracking virus outbreaks in the twenty-first century, Nat. Microbiol. **4**, 10 (2019).

[49] J. Hellewell, S. Abbott, A. Gimma, N. Bosse, C. Jarvis, T. Russell, J. Munday, A. Kucharski, W. Edmunds, CMMID nCoV working group, S. Funk, and R. Eggo, Feasibility of controlling 2019-ncov outbreaks by isolation of cases and contacts, online preprint (2020).

[50] M. E. J. Newman, *Computational Physics* (CreateSpace Independent Publishing Platform, 2012) p. 562.