medRxiv preprint doi: https://doi.org/10.1101/2020.02.10.20021725; this version posted February 12, 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 The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.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: the importance of contact tracing when predicting epidemics

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 (Qc), 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 (Qc), Canada G1V 0A6*

The basic reproductive number — R_0 — is one of the most common and most commonly misapplied numbers in public health. Nevertheless, estimating R_0 for every transmissible pathogen, emerging or endemic, remains a priority for epidemiologists the world over. 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. Here, we show how predicting outbreak size requires both an estimate of R_0 and an estimate of the heterogeneity in the number of secondary infections. To facilitate rapid determination of outbreak risk, we propose a reformulation of a classic result from random network theory 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 how this framework is robust in the face of the typically limited amount of data for emerging pathogens. Lastly, we demonstrate that without data on the heterogeneity in secondary infections for emerging pathogens like 2019-nCoV, the uncertainty in outbreak size ranges dramatically, in the case of 2019-nCoV from 5-40% of susceptible individuals. 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, individuals infected with influenza typically passed on the virus to between 1 and 2 of their social contacts [1]. The same was true for those infected with Ebola virus during the 2014 outbreak in West Africa [2, 3]. Nevertheless, Ebola virus disease infected a tenth of one percent of the number of individuals believed to have been infected by the 1918 Influenza virus [4, 5]. 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 [6], 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 via contact tracing is critically important for quantifying the transmission risk of novel pathogens and further show how a lack of publicly available contact tracing data on cases of novel coronavirus (2019-nCoV) prevents the global public health community from determining the true pandemic risk of this novel virus.

The basic reproduction number of an epidemic, R_0, is the expected number of secondary cases [7], or infections, produced by a primary case over the course of their infectious period in a completely susceptible population [8]. It is a simple metric that is commonly used to describe the transmissibilty of emerging and endemic pathogens [9]. If $R_0 = 2$, one case turns to two, on average, and two turn to four as the epidemic grows. And if $R_0 < 1$, the epidemic will die out. However, we are seldom concerned with epidemics that emerge and quickly die out. To observe an epidemic requires some level of sustained transmission, i.e., $R_0 > 1$. Almost 100 years ago, pioneering work from Kermack and McKendrick [10–12] first demonstrated how to estimate the final size of an epidemic for a pathogen with $R_0 > 1$. Specifically, they consider a scenario such that:

1. the disease results in complete immunity or death,
2. all individuals are equally susceptible,
3. the disease is transmitted in a closed population,
4. contacts occur according to the law of mass-action,
5. 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 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 and small variations in R_0 can lead to vastly different total case counts. Unfortunately, the equation relating R_0

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 February 12, 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 .

to final outbreak size from Kermack and McKendrick is only valid when all the above assumptions hold, which in practice is rarely the case. In fact, seemingly trivial violations of the above assumptions can lead to vastly different outbreak sizes even when R_0 is held constant.

As a result, relying on R_0 alone is often misleading when comparing different pathogens or outbreaks of the same pathogen in different settings [13–15]. This is especially critical considering that most outbreaks are not shaped by the "average" individuals but rather by a minority of superspreading events [13, 16]. For example, public health officials are currently trying determine whether 2019-nCoV can be contained or whether it will cause a pandemic [17]. Despite numerous estimates of 2019-nCoV's R_0 (ranging from 1.5 to just under 4 [18–23]), as we discussed above in the comparison of Ebola virus disease and 1918 influenza, whether—like SARS—2019-nCoV can be contained or—like 2009 H1N1—2019-nCoV will cause a pandemic depends largely on the heterogeneity in the number of secondary infections. Indeed, recent work modeling the effectiveness of contact tracing and isolation on preventing 2019-nCoV concluded that the probability of containing the outbreak was significantly lower if 2019-nCoV heterogeneity in secondary infections was low, like influenza, as compared to higher, like SARS [24]. 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. ANALYSIS FROM NETWORK THEORY

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. [25] and define $G_0(x)$ as the probability generating function (PGF) of the degree distribution $\{p_k\}$ (number of contacts), i.e.

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

When following a random edge, we define the *excess* degree as the number of other edges around that node. 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

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

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 written as

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

because it contains all nodes except those with no edges leading to the GCC. This solution is exact in the limit of large population size.

III. REINTERPRETATION AND 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 a polynomial of high order.

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 underlying variance, κ_3 is related to the skewness and κ_4 to the kurtosis of the full distribution. By definition, a PGF $G(x)$ of a random variable is linked to $K(y)$ through $G(x) = \exp\left[K\left(\ln x\right)\right]$. 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.

medRxiv preprint doi: https://doi.org/10.1101/2020.02.10.20021725; this version posted February 12, 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 .

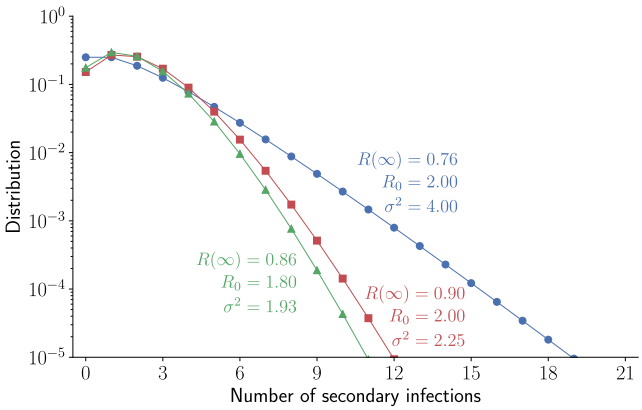

FIG. 1. Illustration of the absence of a direct relationship between R_0 and the final outbreak size, $R(\infty)$, using three different distributions for the number of secondary infections per infected individual. All three distributions are negative binomials [13] with distinct average, R_0, and variance, σ^2. $R(\infty)$ is computed by solving Eqs. (4) and (5).

A. Analysis of cumulants and derivation of Kermack-McKendrick

Using the cumulants κ_n, we can re-write $G_1(x)$ as

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

which is obviously not as practical as a simple polynomial when solving for $u = G_1^C(u)$. However, in some cases, cumulants can have significant advantage over the polynomial representation. For instance, the cumulants of a sum of independent random variables are simply the sum of the cumulants — convenient when dealing with diseases with multiple modes of transmission.

We can quickly derive Kermack and McKendrick's result from the previous generating function since their solution involves a well-mixed population, which is equivalent to a Poisson distribution of secondary infections in our framework. $G_1^C(x)$ is convenient for a Poisson distribution because its cumulants $\kappa_n = R_0\ \forall n$. Moreover, since $G_0(x) = G_1(x)$ in the Poisson case, we know that the final proportion of susceptible individuals is directly given by $u = G_1^C(u)$, or

$$u = \exp\left[\sum_{n=1}^{\infty} \frac{1}{n!} R_0 (\ln u)^n\right] = \exp\left[R_0 (u-1)\right] \ .$$
$$\hookrightarrow R(\infty) = 1 - \exp\left[R_0 (u-1)\right] \ . \tag{7}$$

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

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

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

The solution to $u = G_1^C(u)$ gives the probability that every infection caused by patient zero fails to generate an epidemic. Importantly, Eq. (8) has an alternating nature because $\ln x$ is negative for $x < 1$ such that its n-th power is positive when n is even and negative when n is odd.

This observation, that the moments of Eq. (8) alternate, 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 (see Fig. 1). 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.

Clearly, 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 more easily invade a population and to reach a larger final outbreak size than a disease with a higher R_0. We can investigate this conclusion further using a simple example of normally distributed secondary infections.

B. Normal distributions and the impact of variance

A second useful application of the cumulants formulation involves diseases with a large reproductive number R_0 whose distribution of secondary infections can be convincingly modeled by a normal distribution. The raw moments of a normal distribution are quite complicated, but the cumulants are simple: κ_1 is equal to the mean R_0, κ_2 is equal to the variance σ^2, and all other cumulants are 0. We can thus write

$$G_1^C(x) = \exp\left[R_0 \ln x + \frac{1}{2}\sigma^2 (\ln x)^2\right] = x^{R_0 + \frac{\sigma^2}{2}\ln x} \tag{9}$$

and solving for $u = G_1^C(u)$ yields

$$u = \exp\left[-\frac{2}{\sigma^2}(R_0 - 1)\right] \ . \tag{10}$$

This equation can then be used for direct comparison of the probability of invasion of two different diseases with normal distributions of secondary infections. Given a transmission event from patient zero to a susceptible individual, disease B will be more likely to invade the population than disease A if

$$\frac{\sigma_B^2}{\sigma_A^2} < \frac{R_{0,A} - 1}{R_{0,B} - 1} \ . \tag{11}$$

For example, a disease with half the basic reproductive number of another will still be more likely to invade a population and lead to a larger outbreak if its variance is only slightly less than half the variance of the other disease.

medRxiv preprint doi: https://doi.org/10.1101/2020.02.10.20021725; this version posted February 12, 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 .

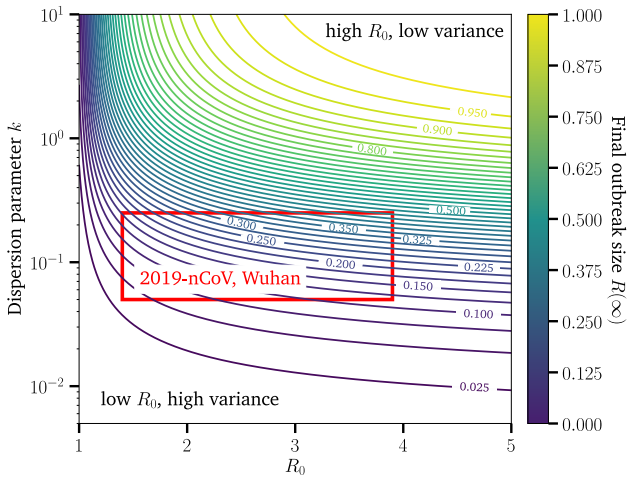

FIG. 2. Final size of outbreaks with different R_0 and distributions of secondary cases. We use a negative binomial distribution of secondary cases and scan realistic range of parameters. Most importantly, the range of parameters corresponding to the current outbreak of 2019-nCoV is highlighted by a red box. The range of potential R0 comes from a 95% confidence interval using a early data and a classic deterministic models [18, 19]. The range of dispersion parameter k comes from analogy with severe acute respiratory syndrome [13, 24]. Most importantly, with fixed average, the dispersion parameter is inversely proportional to the variance of the underlying distribution of second cases.

IV. DISCUSSION

From re-emerging pathogens like yellow fever and measles to emerging threats like MERS-CoV and Ebola, the World Health Organization monitored 119 different infectious disease outbreaks in 2019 alone [26]. For each of these outbreaks, predicting both the epidemic potential and the most likely number of cases is critically important for responding efficiently and effectively. This need for rapid situation awareness is why R_0 is so widely used in public health. However, our main analysis, and Eq. (11) in particular, show 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.

However, 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 [27], metapopulation structure [28, 29], mobility [30, 31], adaptive behavior [32, 33], higher-order contact structure [34, 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.

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 [40–42] and structure of secondary cases [43–45]. For example, methods exist to reconstruct transmission trees for sampled sequences using simple mutational models to construct a likelihood for a specific transmission tree [46, 47] and translate coalescent rates into key epidemiological parameters [48–50]. 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 [51–53].

Third, and most often the first available real time information on novel pathogens, are data related to similar past outbreaks. For example, in Fig. 2 we make a range of predictions for the final size of 2019-nCoV in Wuhan based on R_0 estimates from early cases [18, 19] and for the underlying distribution of secondary cases by analogy with the severe acute respiratory syndrome in similar population [13]. Based on this uncertainty, we obtain a range of final outbreak size (given as a fraction of the total susceptible population) between 5% and 40%. Critically, this large range stems from uncertainty in the heterogeneity of secondary infections. If the heterogeneity is large, sustained transmission is mostly maintained by so-called "super-spreading" events, then the outbreak is both more likely to end stochastically, less likely to spread extensively, as well as easier to manage with contact tracing, screening and infection control [24]. With less heterogeneity, the outbreak almost certainly cannot be contained and we must prepare for a pandemic of 2019-nCoV [54, 55]. Clearly, we are in dire need of contact tracing data and/or high-resolution pathogen genome surveillance for this outbreak.

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 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.

medRxiv preprint doi: https://doi.org/10.1101/2020.02.10.20021725; this version posted February 12, 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 .

[1] 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).

[2] 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**, 1481–1495 (2014).

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

[4] C. E. Mills, J. M. Robins, and M. Lipsitch, "Transmissibility of 1918 pandemic influenza," Nature **432**, 904–906 (2004).

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

[6] 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**, 7588–7593 (2007).

[7] A technical but important caveat: We use the word "case" in a generic sense to represent any infection, even if too mild to meet the clinical case definition.

[8] 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.

[9] 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).

[10] W. O. Kermack and A. G. McKendrick, "A Contribution to the Mathematical Theory of Epidemics," Proc. R. Soc. A **115**, 700–721 (1927).

[11] W. O. Kermack and A. G. McKendrick, "Contributions to the Mathematical Theory of Epidemics. II. The Problem of Endemicity," Proc. R. Soc. A **138**, 55–83 (1932).

[12] 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**, 94–122 (1933).

[13] 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**, 355–359 (2005).

[14] S. Bansal, B. T. Grenfell, and L. A. Meyers, "When individual behaviour matters: homogeneous and network models in epidemiology," J. R. Soc. Interface **4**, 879–891 (2007).

[15] 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.

[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**, 71–81 (2005).

[17] J. T. Wu, K. Leung, and G. M. Leung, "Nowcasting and forecasting the potential domestic and international spread of the 2019-nCoV outbreak originating in Wuhan, China: a modelling study," Lancet (2020), 10.1016/S0140-6736(20)30260-9.

[18] 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. (2020), 10.1056/NEJMoa2001316.

[19] J. M. Read, J. R. E. Bridgen, D. A. T. Cummings, A. Ho, and C. P. Jewell, "Novel coronavirus 2019-nCoV: early estimation of epidemiological parameters and epidemic predictions," medRxiv (2020), 10.1101/2020.01.23.20018549.

[20] T. Liu, J. Hu, M. Kang, L. Lin, H. Zhong, J. Xiao, G. He, T. Song, Q. Huang, Z. Rong, A. Deng, W. Zeng, X. Tan, S. Zeng, Z. Zhu, J. Li, D. Wan, J. Lu, H. Deng, J. He, and W. Ma, "Transmission dynamics of 2019 novel coronavirus (2019-nCoV)," bioRxiv (2020), 10.1101/2020.01.25.919787.

[21] C. Huang, Y. Wang, X. Li, L. Ren, J. Zhao, Y. Hu, L. Zhang, G. Fan, J. Xu, X. Gu, Z. Cheng, T. Yu, J. Xia, Y. Wei, W. Wu, X. Xie, W. Yin, H. Li, M. Liu, Y. Xiao, H. Gao, L. Guo, J. Xie, G. Wang, R. Jiang, Z. Gao, Q. Jin, J. Wang, and B. Cao, "Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China," Lancet (2020), 10.1016/S0140-6736(20)30183-5.

[22] J. Riou and C. L. Althaus, "Pattern of early human-to-human transmission of Wuhan 2019-nCoV," bioRxiv (2020), 10.1101/2020.01.23.917351.

[23] M. Majumder and K. D. Mandl, "Early Transmissibility Assessment of a Novel Coronavirus in Wuhan, China," SSRN Electron. J. (2020), 10.2139/ssrn.3524675.

[24] J Hellewell, S Abbott, A Gimma, NI Bosse, CI Jarvis, TW Russell, JD Munday, AJ Kucharski, WJ Edmunds, CMMID nCoV working group, S Funk, and RM Eggo, "Feasibility of controlling 2019-ncov outbreaks by isolation of cases and contacts," online preprint (2020).

[25] 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).

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

[27] Y. Moreno, R. Pastor-Satorras, and A. Vespignani, "Epidemic outbreaks in complex heterogeneous networks," Eur. Phys. J. B **26**, 521–529 (2002).

[28] V. Colizza and A. Vespignani, "Epidemic modeling in metapopulation systems with heterogeneous coupling pattern: Theory and simulations," J. Theor. Biol. **251**, 450–467 (2008).

[29] D. J. Watts, R. Muhamad, D. C. Medina, and P. S. Dodds, "Multiscale, resurgent epidemics in a hierarchical metapopulation model," Proc. Natl. Acad. Sci. USA **102**, 11157–11162 (2005).

[30] V. Colizza, A. Barrat, M. Barthélemy, and A. Vespignani, "The role of the airline transportation network in the prediction and predictability of global epidemics," Proc. Natl. Acad. Sci. USA **103**, 2015–2020 (2006).

[31] 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**, 11887–11892 (2015).

[32] S. V. Scarpino, A. Allard, and L. Hébert-Dufresne, "The effect of a prudent adaptive behaviour on disease transmission," Nat. Phys. **12**, 1042–1046 (2016).

[33] B. M. Althouse and L. Hébert-Dufresne, "Epidemic cycles driven by host behaviour," J. R. Soc. Interface **11**, 20140575

medRxiv preprint doi: https://doi.org/10.1101/2020.02.10.20021725; this version posted February 12, 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 .

(2014).

[34] L. Hébert-Dufresne, P.-A. Noël, V. Marceau, A. Allard, and L. J. Dubé, "Propagation dynamics on networks featuring complex topologies," Phys. Rev. E **82**, 036115 (2010).

[35] E. M. Volz, J. C. Miller, A. Galvani, and L. Ancel Meyers, "Effects of Heterogeneous and Clustered Contact Patterns on Infectious Disease Dynamics," PLOS Comput. Biol. **7**, e1002042 (2011).

[36] L. Hébert-Dufresne and B. M. Althouse, "Complex dynamics of synergistic coinfections on realistically clustered networks," Proc. Natl. Acad. Sci. USA **112**, 10551–10556 (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**, 1302–1304 (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**, 1122–1125 (2009).

[41] S. K. Gire, A. Goba, K. G. Andersen, R. S. G. Sealfon, D. J. Park, L. Kanneh, S. Jalloh, M. Momoh, M. Fullah, G. Dudas, S. Wohl, L. M. Moses, N. L. Yozwiak, S. Winnicki, C. B. Matranga, C. M. Malboeuf, J. Qu, A. D. Gladden, S. F. Schaffner, X. Yang, P.-P. Jiang, M. Nekoui, A. Colubri, M. R. Coomber, M. Fonnie, A. Moigboi, M. Gbakie, F. K. Kamara, V. Tucker, E. Konuwa, S. Saffa, J. Sellu, A. A. Jalloh, A. Kovoma, J. Koninga, I. Mustapha, K. Kargbo, M. Foday, M. Yillah, F. Kanneh, W. Robert, J. L. B. Massally, S. B. Chapman, J. Bochicchio, C. Murphy, C. Nusbaum, S. Young, B. W. Birren, D. S. Grant, J. S. Scheiffelin, E. S. Lander, C. Happi, S. M. Gevao, A. Gnirke, A. Rambaut, R. F. Garry, S. H. Khan, and P. C. Sabeti, "Genomic surveillance elucidates Ebola virus origin and transmission during the 2014 outbreak," Science **345**, 1369–1372 (2014).

[42] C. Drosten, D. Muth, V. M. Corman, R. Hussain, M. Al Masri, W. HajOmar, O. Landt, A. Assiri, I. Eckerle, A. Al Shangiti, J. A. Al-Tawfiq, A. Albarrak, A. Zumla, A. Rambaut, and Z. A. Memish, "An Observational, Laboratory-Based Study of Outbreaks of Middle East Respiratory Syndrome Coronavirus in Jeddah and Riyadh, Kingdom of Saudi Arabia, 2014," Clin. Infect. Dis. **60**, 369–377 (2015).

[43] 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 Jeffrey P. Townsend, "Epidemiological and Viral Genomic Sequence Analysis of the 2014 Ebola Outbreak Reveals Clustered Transmission," Clin. Infect. Dis. **60**, 1079–1082 (2015).

[44] A. Arias, S. J. Watson, D. Asogun, E. A. Tobin, J. Lu, M. V. T. Phan, U. Jah, R. E. G. Wadoum, L. Meredith, L. Thorne, S. Caddy, A. Tarawalie, P. Langat, G. Dudas, N. R. Faria, S. Dellicour, A. Kamara, B. Kargbo, B. O. Kamara, S. Gevao, D. Cooper, M. Newport, P. Horby, J. Dunning, F. Sahr, T. Brooks, A. J. H. Simpson, E. Groppelli, G. Liu, N. Mulakken, K. Rhodes, J. Akpablie, Z. Yoti, M. Lamunu, E. Vitto, P. Otim, C. Owilli, I. Boateng, L. Okoror, E. Omomoh, J. Oyakhilome, R. Omiunu, I. Yemisis, D. Adomeh, S. Ehikhiametalor, P. Akhilomen, C. Aire, A. Kurth, N. Cook, J. Baumann, M. Gabriel, R. Wölfel, A. Di Caro, M. W. Carroll, S. Günther, J. Redd, D. Naidoo, O. G. Pybus, A. Rambaut, P. Kellam, I. Goodfellow, and M. Cotten, "Rapid outbreak sequencing of Ebola virus in Sierra Leone identifies transmission chains linked to sporadic cases," Virus Evol. **2**, vew016 (2016).

[45] D. J. Park, G. Dudas, S. Wohl, A. Goba, S. L. M. Whitmer, K. G. Andersen, R. S. Sealfon, J. T. Ladner, J. R. Kugelman, C. B. Matranga, S. M. Winnicki, J. Qu, S. K. Gire, A. Gladden-Young, S. Jalloh, D. Nosamiefan, N. L. Yozwiak, L. M. Moses, P.-P. Jiang, A. E. Lin, S. F. Schaffner, B. Bird, J. Towner, M. Mamoh, M. Gbakie, L. Kanneh, D. Kargbo, J. L. B. Massally, F. K. Kamara, E. Konuwa, J. Sellu, A. A. Jalloh, I. Mustapha, M. Foday, M. Yillah, B. R. Erickson, T. Sealy, D. Blau, C. Paddock, A. Brault, B. Amman, J. Basile, S. Bearden, J. Belser, E. Bergeron, S. Campbell, A. Chakrabarti, K. Dodd, M. Flint, A. Gibbons, C. Goodman, J. Klena, L. McMullan, L. Morgan, B. Russell, J. Salzer, A. Sanchez, D. Wang, I. Jungreis, C. Tomkins-Tinch, A. Kislyuk, M. F. Lin, S. Chapman, B. MacInnis, A. Matthews, J. Bochicchio, L. E. Hensley, J. H. Kuhn, C. Nusbaum, J. S. Schieffelin, B. W. Birren, M. Forget, S. T. Nichol, G. F. Palacios, D. Ndiaye, C. Happi, S. M. Gevao, M. A. Vandi, B. Kargbo, E. C. Holmes, T. Bedford, A. Gnirke, U. Ströher, A. Rambaut, R. F. Garry, and P. C. Sabeti, "Ebola Virus Epidemiology, Transmission, and Evolution during Seven Months in Sierra Leone," Cell **161**, 1516–1526 (2015).

[46] 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).

[47] 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).

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

[49] 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).

[50] T. Stadler, D. Kuhnert, S. Bonhoeffer, and A. J. Drummond, "Birth-death skyline plot reveals temporal changes of epidemic spread in HIV and hepatitis C virus (HCV)," Proc. Natl. Acad. Sci. USA **110**, 228–233 (2013).

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

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

[53] 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**, 10–19 (2019).

[54] V. J. Munster, M. Koopmans, N. van Doremalen, D. van Riel, and E. de Wit, "A Novel Coronavirus Emerging in China — Key Questions for Impact Assessment," N. Engl. J. Med. (2020), 10.1056/NEJMp2000929.

[55] N. Imai, A. Cori, I. Dorigatti, M. Baguelin, C. A. Donnelly, S. Riley, and N. M. Ferguson, "Report 3: Transmissibility of 2019-nCoV," online preprint (2020).