Superspreaders and High Variance Infectious Diseases

Yaron Oz[1], Ittai Rubinstein[2], and Muli Safra[2]
[1] *Raymond and Beverly Sackler School of Physics and Astronomy, Tel-Aviv University, Tel-Aviv 69978, Israel*
[2] *Blavatnik School of Computer Science, Tel-Aviv University, Tel-Aviv 69978, Israel*
(Dated: September 6, 2020)

A well-known characteristic of pandemics such as COVID-19 is the high level of transmission heterogeneity in the infection spread: not all infected individuals spread the disease at the same rate and some individuals (superspreaders) are responsible for most of the infections. To quantify this phenomenon requires the analysis of the effect of the variance and higher moments of the infection distribution. Working in the framework of stochastic branching processes, we derive an approximate analytical formula for the probability of an outbreak in the high variance regime of the infection distribution, verify it numerically and analyze its regime of validity in various examples. We show that it is possible for an outbreak not to occur in the high variance regime even when the basic reproduction number R_0 is larger than one and discuss the implications of our results for COVID-19 and other pandemics.

PACS numbers: 87.10.+e

I. INTRODUCTION

The classic SIR models provide an epidemiology framework for studying the spread of a disease [1]. The basic reproduction number R_0 in these models is the mean value of secondary infections caused by an infected individual. It determines the threshold $R_0 > 1$ for an outbreak. Alternatively, it determines the fraction of the population that will be infected before herd immunity is reached. In view of the importance of this parameter, major measures (such as lockdowns) are taken in order to reduce the value of R_0. The estimation for the COVID-19 pandemic, for example, is $R_0 \sim 2-3$.

The structure underlying the epidemic spreading is that of a complex heterogeneous network, where a small number of the nodes act as hubs while the majority of nodes have few contacts (for a review and references therein see [2]). Indeed, not all people cause a similar number of secondary infections and there is clear empirical evidence for high levels of transmission heterogeneity in the infection spread (see e.g. [3–5]). The analysis in [4] for the COVID-19 pandemic suggests that between 5% to 10% of infected individuals are responsible for 80% of secondary infections. This may be due to differences in the number of contacts, in protective equipment, in levels of hygiene, in time of diagnosis or biological effects such as tendency to cough and sneeze.

Individuals with high secondary infection rate are commonly referred to as *superspreaders*. This is encoded in the degree distribution of the epidemic spread network corresponding to the *infection distribution*. While homogeneous random networks such as the Erdos-Renyi model exhibit a statistical homogeneity of the nodes and the degree distribution is peaked around the average value, heterogeneous networks such as the scale free models reveal a power law structure of the degree distribution and nodes with very large degree.

The infection distribution is taken not over a random individual, but rather over a random infected individual, i.e. it is weighted according to the *a priori* probability of each individual to be infected. For instance, an individual in contact with many people has a higher likelihood both to be infected and to infect others and this is reflected in the degree of the corresponding node in the network.

Studying the phenomenon of superspreaders, which seems to follow the Pareto-type Principle [6], as well as its implications on the spread of the disease is crucial when devising and implementing control policies [6, 7]. In order to analyze the impact of the superspreaders on the epidemic spread we have to consider the effect of the variance and the higher moments of the infection distribution. The main goal of this paper, is to study a question of utmost importance when facing pandemics such as COVID-19, namely:"what is the probability that a disease will disappear without a major outbreak?"

An outbreak is often referred to as a sudden rise in the number of infected individuals. In this paper, however, we define an outbreak with reference to the total fraction of infected individuals in the long term and not at any specific point in time. Thus, we consider that *an outbreak has not occurred* if the disease has disappeared with a negligible herd immunity. Note, that we will analyse the natural evolution of the disease irrespective of the measures—social and others—taken to reduce R_0.

We will work in the framework of Galton-Watson branching processes (for a review see e.g. [8]), and use it to predict the probability of an outbreak as a function of the infection distribution, that is the probability distribution for an individual to infect a given number of people. We derive an approximate analytical formula for the probability of an outbreak in the high variance regime of the infection distribution, verify it numerically in various examples, compare it to COVID-2 data and discuss its implications for the COVID-19 pandemic. In particular, we will show that it is possible for an outbreak not to occur in the high variance regime even when the basic reproduction number R_0 is larger than one. This phenomenon has been observed in numerical simulations

[9].

II. THE HIGH VARIANCE REGIME

The infection distribution specifies, for each natural number k, the probability of an infected individual to infect k others. We denote by R_0 and V the mean and variance of the number of people infected. When $R_0 < 1$, it is well established that the disease would disappear on its own, while when $R_0 - 1$ is not small compared to the variance V, one can use deterministic models such as SIR that provide an accurate description.

Let us thus focus on the high variance regime:

$$0 < R_0 - 1 \ll V \ . \tag{II.1}$$

Our main result can be stated as follows. The probability that a disease will disappear without herd immunity is:

$$\mathsf{Pr} = \gamma^n, \tag{II.2}$$

where n is the current number of infected individuals and γ in the regime (II.1) can be approximated as:

$$\gamma \approx 1 - \mathcal{Q} \ , \tag{II.3}$$

where

$$\mathcal{Q} = \frac{2(R_0 - 1)}{R_0^2 + V - R_0} \ . \tag{II.4}$$

Below, the corrections to the approximate formula (II.3) and (II.4) are bounded by higher powers of the ratio $\mathcal{Q}$ as well as the higher moments of the infection distribution.

In section III we formulate the main result precisely and prove it. However, before delving into the proof let us consider some of its qualitative implications, compare it to pandemic data and numerically verify its accuracy. First, the larger the variance V compared to $R_0 - 1$, the higher the probability for the disease to disappear before herd immunity is reached. Thus, the fate of the disease does not depend only on $R_0 - 1$. Second, the fewer infected individuals, the higher the probability for the disease to disappear and, consequently, the less stringent the pandemic measures that must be taken, even when $R_0 > 1$. Third, the effective dependence on the variance is $\frac{V}{n}$.

Let us numerically compare our approximate analytical formula to the exact γ for the re-scaled infection distribution of COVID-2 [3]. The latter is based on fitting pandemic data to a distribution obtained by sampling a Poisson distribution whose mean is sampled from a Gamma distribution, which we will call Gamma-Poisson distribution. Since R_0 of COVID-2 is high, we define the infection distribution for lower values of R_0 by re-scaling the original one, that is, we fix the shape of the distribution that is determined by a parameter k and re-scale the parameter θ that determines the scale of the distribution.

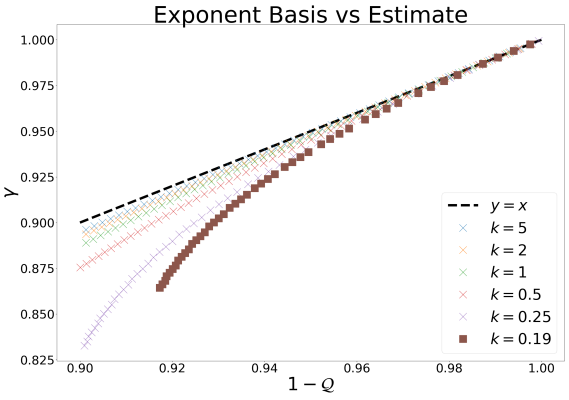

FIG. 1: Our approximate formula $1 - \mathcal{Q}$ vs. the exact Galton-Watson coefficient γ for the re-scaled infection distribution of COVID-2 [3]. The latter fits data to a distribution obtained by sampling a Poisson distribution with mean being sampled from a Gamma distribution. The parameter $k \approx 0.19$ for COVID-2 controls the shape of the Gamma distribution, while the parameter θ controls its scale. We constructed our data by fixing k and scale θ to give different values of $R_0 = k\theta$. We reach $R_0 \approx 1.6$ for the lower value of k.

The results are depicted in figure 1 and, as expected, we see that lower R_0 implies better accuracy.

We use our formula to estimate the probability to avoid an outbreak for the COVID-2 and COVID-19 pandemics as a function of R_0 and the number of infected n—requiring an estimate of the ratio $\frac{V}{R_0^2}$. Based on [3] we set $V = 5R_0^2$ for COVID-2. As noted above, the analysis in [4] estimates that the p_h value (the percentage of the infected population responsible for 80% of all secondary infections) is around $5\% - 10\%$. Assuming Gamma-Poisson distribution, $p_h = 10\%$ implies $k = 0.1$ and $V = 10R_0^2$, and lower p_h values correspond to even higher variance. We plot the results in figures 2 and 3: for given values of R_0 and n, the higher the variance the higher the probability of avoiding an outbreak.

While it is clear that γ cannot be determined precisely by R_0 and V alone and the information about the higher moments of the infection distribution is necessary, our numerical analysis reveals that for certain distributions that are often being employed for real world pandemics the accuracy of (II.3) is mostly determined by the value of $R_0 - 1$ as depicted in figure 4 and figure 5.

In figure 4, the Poisson distributions has λ values in the range 1.0 to 1.1, the geometric distributions has p values in the range 0.43 to 0.5, the Poisson$10x$ distribution is obtained by selecting a Poisson distribution with $10 \leq \lambda \leq 20$ value with probability 10% or the zero distribution with probability 90%, and the Truncated Power Law distributions has a cut-off at 100 with powers ranging from 2.1 to 2.375. In figure 5 we consider the ratio between the logarithms since this determines the ratio between the values of n that would give a specific probability to avoid an outbreak.

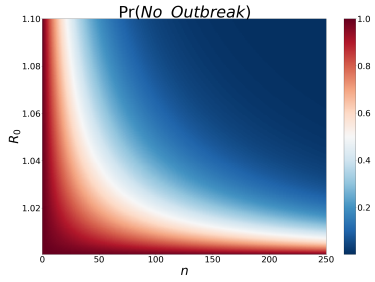

FIG. 2: The probability to avoid an outbreak when $V = 5R_0^2$ (COVID-2) as a function of the basic reproduction number and the number of infected individuals.

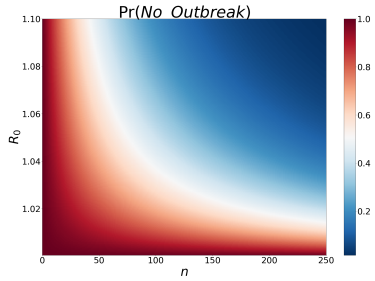

FIG. 3: The probability to avoid an outbreak when $V = 10R_0^2$ (COVID-19) as a function of the basic reproduction number and the number of infected individuals.

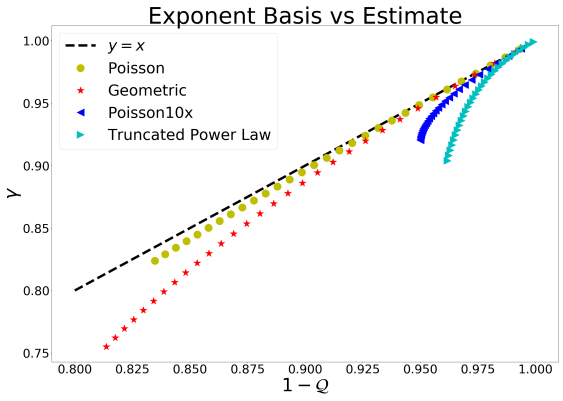

FIG. 4: The exact Galton-Watson coefficient γ vs. our approximate formula $1 - \mathcal{Q}$ for various infection distributions. The line $y = x$ corresponds to $\gamma = 1 - \mathcal{Q}$. The accuracy of the formula depends mostly upon the value of R_0, which explains the different deviations of the distributions from the $y = x$ line: R_0 for the Poisson, Geometric, Poisson10x and the truncated power law distributions are in the ranges $[1, 1.1]$, $[1, 1.32]$, $[1, 2]$, $[1, 1.72]$, respectively.

III. FORMAL STATEMENTS AND PROOFS

We define the infection distribution by a sequence of real variables a_k, where a_k is the probability that a carrier infects k individuals and is removed. The normalization condition is:

$$\sum_k a_k = 1. \tag{III.1}$$

Denote by M_i the ith moment of the infection distribution:

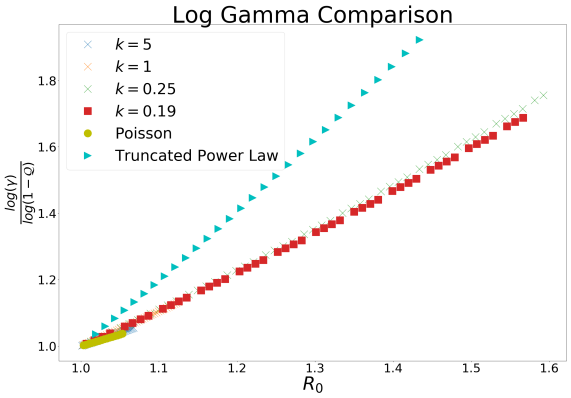

FIG. 5: A comparison of our formula to the exact value of γ for various distributions as a function of R_0, and the larger R_0 the larger the deviation. We tested our results for several distributions: Gamma-Poisson distributions for COVID-2 [3], Poisson and Truncated Power Law distributions.

$$M_i = \sum_k a_k k^i, \quad M_1 = R_0, \quad M_2 = V + R_0^2 \ , \tag{III.2}$$

and the quantity η by:

$$\eta = \frac{1}{R_0^2 + V - R_0} \sum_{i \geq 3} \left(\frac{3\mathcal{Q}}{2} \right)^{i-2} \frac{M_i}{i!} \ . \tag{III.3}$$

Let $p(x)$ be the polynomial

$$p(x) = \sum_k a_k x^k - x \ . \tag{III.4}$$

The proof of our result (II.2) and (II.3) consists of proving three statements:

- $\Pr = \gamma^n$ where $0 \leq \gamma < 1$ is a root of $p(x)$.
- $p(x)$ has a root within a small neighborhood of $1 - \mathcal{Q}$.
- $p(x)$ has at most a single root in $[0, 1)$.

Let us prove the following claims:

Claim III.1 (Galton-Watson analysis)**.** *If $R_0 > 1$ then the probability that the disease will disappear without herd immunity is γ^n where γ satisfies $p(\gamma) = 0$ and $0 \leq \gamma < 1$.*

Claim III.2 (Approximate Formula)**.** *There exist c and $\eta_0 > 0$ s.t. if $\eta < \eta_0$ then $p(x)$ has a root within the interval*

$$[1 - (1 - c\eta)\mathcal{Q}, 1 - (1 + c\eta)\mathcal{Q}]$$

Claim III.3 (Single Root)**.** *$p(x)$ has exactly one root in the interval $[0, 1)$.*

Combining the above assertions, we see that if:

- $R_0 > 1$ (condition for Claim $III.1$)
- $\eta < \eta_0$ (condition for Claim $III.2$)
- $\mathcal{Q} < \frac{1}{(1+c\eta_0)}$ (the root in Claim $III.2$ is positive)

then we arrive at our main result:

$$1 - \gamma \in [1 - c\eta, 1 + c\eta]\mathcal{Q} \ . \tag{III.5}$$

Claim $III.1$ is a standard analysis of Galton-Watson processes [8], which we will now briefly review for completeness.

One views the number of sick individuals as a Markov process, where at each point we pick a sick individual, add the number of people infected by him and remove him. As above, a_k is the transition probability from a state with n sick people to a state with k added infected people and one removed:

$$n \to n + k - 1 \ . \tag{III.6}$$

Denote by $f(n)$ the probability that no major outbreak will occur at any time $t > 0$ if we have at $t = 0$ n infected people, and define $\gamma = f(1) \in [0, 1]$.

For the disease to die out, every branch that begins from one of the n infected individuals at $t = 0$ should disappear. Since we neglect the interaction between the infected individuals, these are independent random variables and $f(n) = f(1)^n = \gamma^n$.

Using time independence and the total probability, one gets the recursion relation:

$$f(n) = \sum_k a_k f(n + k - 1) \ . \tag{III.7}$$

Setting $n = 1$ in (III.7) we have:

$$\sum_k a_k \gamma^k - \gamma = 0 \ , \tag{III.8}$$

that is, γ is a root of the polynomial p(x) (III.4).

Finally, in order to complete the proof of Claim $III.1$, we have to show that $\gamma \neq 1$. This is not surprising, as we are dealing with the $R_0 > 1$ regime and setting $\gamma = 1$ would make the probability of an outbreak $1 - 1^n = 0$ regardless of the number of infected at $t = 0$. In order to prove the claim, we have to show that the probability of an outbreak converges to 1 as $n \to \infty$, but this is easy to see (for instance, by applying Chebyshev's inequality on the probability that n sick will infect less than $\frac{R_0+1}{2}n$ individuals).

Consider next Claim $III.2$. It is convenient to denote $\gamma = 1 + \delta$ and analyze the roots of $p(x)$:

$$p(1 + \delta) = \sum_k a_k (1 + \delta)^k - (1 + \delta) = 0 \ . \tag{III.9}$$

Expanding (III.9) and using (III.1) and (III.2) we get:

$$p(1+\delta) = (R_0-1)\delta + \frac{1}{2}\left(R_0^2 + V - R_0\right)\delta^2 + \ corrections \ . \tag{III.10}$$

From (III.10) we get the approximate formula (II.3) where the corrections are bounded by:

$$corrections \ \leq \sum_{i\geq 3} \frac{M_i}{i!}\delta^i \ . \tag{III.11}$$

Let $\eta_0 = \frac{1}{10}$ and $c = 5$. We are interested in the case where

$$\begin{aligned} \delta &\in [-1 - c\eta, -1 + c\eta]\ \mathcal{Q} \\ &\subseteq -\left[\frac{3\mathcal{Q}}{2}, \frac{\mathcal{Q}}{2}\right] \ . \end{aligned} \tag{III.12}$$

Therefore:

$$\begin{aligned} &p(1 + \delta) = \\ &= (R_0 - 1)\delta + \frac{1}{2}\left(R_0^2 + V - R_0\right)\delta^2 \pm \\ &\quad (R_0^2 + V - R_0)\delta^2\eta \ , \end{aligned} \tag{III.13}$$

where we denote $X = Y \pm Z$ iff $|X - Y| < Z$. It is straightforward to see that when $\delta = -(1 + 5\eta)\mathcal{Q}$ we have $p(1 + \delta) \geq 0$, while when $\delta = -(1 - 5\eta)\mathcal{Q}$, we have $p(1 + \delta) \leq 0$. Combining these results with the Intermediate Value Theorem, we conclude the proof of Claim $III.2$.

In order to prove Claim $III.3$, consider the second derivative of $p(x)$:

$$p''(x) = \frac{d^2p(x)}{dx^2} = \sum_{k\geq 2} k(k - 1)x^{k-2} \ , \tag{III.14}$$

and $p''(x) > 0$ for $x > 0$. Thus, $p(x)$ is convex in $\mathbb{R}^+$, and must have at most two non-negative roots. Using (III.1) we see that $x = 1$ is one of these non-negative roots. Furthermore, $x = 1$ is not a local minimum of $p(x)$, since

$$p'(1) = \sum_k ka_k - 1 = R_0 - 1 > 0 \ , \tag{III.15}$$

and in particular it cannot be the global minimum for $p(x)$ in $x \in \mathbb{R}^+$. This implies that $p(x)$ must have a negative value.

$\forall x > 1$: $p''(x) > 0$ implies that $p'(x) > p'(1) > 0$ and hence $p(x) > 0$. Therefore, p reaches its minimum in the $\mathbb{R}^+$ region at some point b, $0 \leq b < 1$. From the Intermediate Value Theorem, there is a point c, $0 \leq c < b < 1$ such that $p(c) = 0$, and it is clearly unique, concluding our proof.

IV. DISCUSSION AND OUTLOOK

We have carried out an analysis of the stochastic spread of a disease in the high variance regime of the infection distribution. This allowed us to study an important characteristic of the COVID-19 and other pandemics where not all infected individuals spread the disease at the same rate and superspreaders are responsible

for most of the infections. We derived an approximate analytical formula (II.2 and II.3) for the probability to avoid an outbreak in the high variance regime (II.1) and estimated its accuracy numerically and analytically. We found out that $R_0 - 1$ is the main control parameter for the higher moment corrections. Curiously, for all the distributions that we analyzed we found that $\gamma \leq 1 - \mathcal{Q}$, giving us an upper bound on the approximation. We compared the formula to infection distribution data and discussed its implications for the COVID-2 and COVID-19 pandemics.

Our analysis reveals the general coarse-grained structure of the infectious diseases irrespective of the detailed graph or network structure of the disease spread. We studied the natural evolution of the disease under the assumption that the infection and recovery are time-independent random variables. There are several reasons to consider the time dependence of R_0, V and the higher moments, an obvious one being the measures, social and other, taken to reduce them. A less obvious one is related to the time-dependent details of the disease's evolution structure. There can be a major change due to a reduction in the number of superspreaders that are removed, which leads to interesting insights about the disease spread, such as reaching herd immunity faster than previously assumed [10].

Acknowledgements

We would like to thank Nir Kalkstein for valuable discussions on the importance of the high variance to the spread of the disease, as well as Baruch Barzel for comments on the manuscript. The work is supported in part by the Israeli Science Foundation center of excellence. The work is supported in part by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (Grant agreement No. 835152), as well as by ISF 2013/17 and BSF 2016414.

[1] W. O. Kermack and A. G. McKendrick, "A contribution to the mathematical theory of epidemics", Proceedings of the Royal Society A, Volume 115, Issue 772 (1927).
[2] R. Pastor-Satorras, C. Castellano, P. Van Mieghem and A. Vespignani, "Epidemic processes in complex networks", Rev. Mod. Phys. **87**, 925 (2015).
[3] 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).
[4] D. Miller et al.,"Full genome viral sequences inform patterns of SARS-CoV-2 spread into and within Israel", doi:https://doi.org/10.1101/2020.05.21.20104521.
[5] T. Britton, F. Ball, P. Trapman, "A mathematical model reveals the influence of population heterogeneity on her immunity to SARS-Cov-2", Science 10.1126/science.abc6810 (2020).
[6] M. E. J. Woolhouse et al., "Heterogeneities in the transmission of infectious agents: Implications for the design of control programs", Proc. Natl Acad. Sci. USA **94**, 338 (1997).
[7] R. Pastor-Satorras and A. Vespignani, "Immunization of complex networks", Phys. Rev. E. **65**, 036104 (2002).
[8] D. R. Insua, F. Ruggeri, M. P. Wiper, "Bayesian Analysis of Stochastic Process Models", Wiley (2012).
[9] B. Barzel, private communication.
[10] Y. Oz, I. Rubinstein and M. Safra, to appear.