medRxiv preprint doi: https://doi.org/10.1101/2020.09.06.20189290; this version posted September 10, 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 The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.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 .

Heterogeneity and Superspreading Effect on Herd Immunity

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)

We model and calculate the fraction of infected population necessary for herd immunity to occur, taking into account the heterogeneity in infectiousness and susceptibility, as well as the correlation between the two parameters. We show that these cause the reproduction number to decrease with progression, and consequently have a drastic effect on the estimate of the necessary percentage of the population that has to contract the disease for herd immunity to be reached. We discuss the implications to COVID-19 and other pandemics.

PACS numbers: 87.10.+e

I. INTRODUCTION

The COVID-19 pandemic has had a dramatic impact on the world in recent months, setting in motion a huge effort in diverse research disciplines in an attempt to understand the nature of the disease and the dynamics of the virus's spread. A question of utmost importance in this context is how many infected individuals it takes to reach herd immunity, where by herd immunity one means that the virus is unable to find enough susceptible hosts to continue its spread and consequently the disease fades out.

Herd immunity is typically expected to be reached when a large fraction of the population becomes immune to the disease and the effective reproduction number (which quantifies the spread) drops below one. The standard estimate for the necessary fraction for herd immunity to be reached is about 60% of the susceptible individuals. This estimate, however, assumes a homogeneous structure of the epidemic spread network, where both the infectiousness and susceptibility of individuals are assumed to be homogeneously distributed.

It is well recognized, however, that the epidemic spread network is not homogeneous but rather heterogeneous, with distinct people being *infectious* (likely to infect others) and *susceptible* (likely to become infected themselves) to different degrees (for reviews see e.g. [1, 2]). The class of individuals with very high secondary infection rates are referred to as superspreaders [3]. An estimate for the COVID-19 pandemic [4] asserts that between 5% to 10% of the infected individuals cause 80% of the secondary infections.

The reasons that different people are infectious and susceptible to different degrees may be, for instance, increased contact with others that can increase both parameters, or hygiene and better protective equipment that can decrease both. Thus, infectiousness and susceptibility are potentially highly correlated. A correlation between infectiousness and susceptibility can significantly affect our estimate for the percentage of the population that must contract the disease for herd immunity to be reached.

If more infectious people are also more susceptible, then our initial estimates of the basic reproduction number (the mean value of secondary infections caused by an infected individual) will be biased leading us to believe that it is much larger than it really is—we are oversampling the infectiousness of more susceptible people. Deviations in the susceptibility can lead to us seeing an early spike in the number of cases as the susceptible are infected, with a sudden drop later on as the disease spreads to less susceptible populations. Furthermore, if more infectious people are also more susceptible, then they will also be infected and develop natural immunity much sooner.

Our aim in this work is to model and calculate the fraction of infected in the population that gives rise to herd immunity while taking into account the heterogeneity in infectiousness and susceptibility, their correlation and the superspreading effect.

In order to analyze the spread of the disease we assign to each individual a a susceptibility parameter $S(a)$ and an infectiousness parameter $I(a)$ drawn from some probability distributions. $S(a)$ and $I(a)$ quantify how likely a is to be infected and infect others, respectively. The probability that a will transmit the disease to b—should a be infected—is $I(a) \times S(b)$ and this is the effective parameter in our analysis. The product of I and S scales like the inverse of the susceptible population, where typically S scales like its inverse and I is independent of it.

We measure the progress of the disease as a function of the number of individuals who contracted it. This is the natural governing parameter when considering questions of the type: will the disease fade out as a function of the fraction of the population that got infected. We begin the evolution process of the disease at step $n = 0$ with a certain number of infected individuals and increase the number of infected by one at each step. The initial number of infected individuals does not influence the analysis and we will take it to be one. Our calculation is done by taking an expectation value over all possible scenarios of infection.

We model the system for any distribution of I, S, and present a general formula for the behaviour of heterogeneous diseases. We also consider the special case where

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/10.1101/2020.09.06.20189290; this version posted September 10, 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 individual is infectious and susceptible to the same degree, that is, where the distributions from which I and S are drawn are highly correlated. We derive a simple analytical result when the infectiousness and susceptibility parameters are, in particular, chosen from a Gamma distribution with scaling and shape parameters k and θ, respectively—a distribution previously attributed to the infectiousness of COVID-2 [3].

Consider first the general case. Define the average conditional infectiousness $\varphi(s)$:

$$\varphi(s) \stackrel{def}{=} \mathop{\mathbb{E}}_{S(a)=s} [I(a)] \ , \tag{I.1}$$

and the normalized susceptibility distribution at the step n in the evolution of the disease $\rho(s,n)$:

$$\rho(s,n) \stackrel{def}{=} \Pr\left[(S(a)=s) \,|\, a \text{ is healthy after } n \text{ steps} \right] . \tag{I.2}$$

Denote the susceptibility distribution at the beginning by $\rho(s) = \rho(s,0)$. The average conditional infectiousness (I.1) is independent of the number of infected individuals while the susceptibility distribution does depend on it—individuals with higher susceptibility are more likely to be chosen first, hence their rate decreases as the process progresses. We will prove the following general claim about the fraction of the population necessary to reach herd immunity:

Claim I (General Case): *for any δ when*

$$1 - \int \rho(\sigma) exp(-\delta\sigma) d\sigma \tag{I.3}$$

fraction of the population is infected, the effective reproduction number will be reduced by a factor of

$$\frac{\int \varphi(\sigma)\rho(\sigma)\sigma \exp\left(-\delta\sigma\right) d\sigma}{\int \varphi(\sigma)\rho(\sigma)\sigma d\sigma} \ . \tag{I.4}$$

The threshold for herd immunity is when the value of the effective reproduction number is 1.

Consider next the particular case of the Gamma distribution with shape and scale parameters k and θ, respectively. We will prove the following claim:

Claim II (Gamma distribution): *Under the above assumptions, herd immunity will be reached when*

$$\varepsilon = 1 - R^{\frac{-k}{(k+2)}} \tag{I.5}$$

fraction of the population is infected. R is the reproduction number at the beginning of the disease spread and k is the shape (spread) parameter of the Gamma distribution.

Note, that the necessary fraction of the population (I.5) does not depend on the scale parameter of the distribution θ. Substituting the estimates for COVID-19: $R \approx 3$ and $k \approx 0.1$ [5, 6], we get $\varepsilon \approx 5\%$. This result is far more optimistic than the recent estimate [7] that requires about 40% of the population to contract the disease before herd immunity is achieved. It also suggests a possible explanation for the observation that the COVID-19 pandemic appears to be slowing down despite the relatively low number of infections so far [4]. The Gamma distribution has been used in order to model the data of COVID-2 with $k = 0.19$ [3] and in this case we get $\varepsilon \approx 9\%$.

The letter is organized as follows. In section II we will define precisely the effective reproduction number that we will use in our analysis. In section III we will study the dynamics of the disease spread and prove claim I. In section IV we will prove claim II and study numerically claim I for various types of infection and susceptibility distributions. Section V is devoted to a discussion.

II. EFFECTIVE REPRODUCTION NUMBER

Consider the effective value of the reproduction number at step n in the evolution of the pandemic which we will denote by $R(n)$, $R = R(0)$. We define it as the expectation value of secondary infections conditional on the individuals that have been infected. Denote by Λ_n the distribution over the nth individual to be infected (namely, linear in susceptibility at that stage)—we thus have:

$$R(n) \stackrel{def}{=} \mathop{\mathbb{E}}_{a\sim\Lambda_n} [\text{number of people } a \text{ infects}] \ . \tag{II.1}$$

Mathematically we have:

$$\begin{aligned} R(n) &= \mathop{\mathbb{E}}_{a\sim\Lambda_n} \left[I(a) \sum_{b\neq a} S(b) \right] = \\ &= \sum_a \frac{S(a)}{\sum_{a'} S(a')} I(a) \sum_{b\neq a} S(b) = \\ &\simeq \sum_a S(a) I(a) \ , \end{aligned} \tag{II.2}$$

where the summation over a in the last two lines is on the healthy individuals at step $n-1$ and in the approximation at the last line we added the infected individual a to the summation. Note, that since $I \times S$ scales like the inverse of the susceptible population at the nth step $N(n)$, R (II.2) does not scale with it. Using (I.1) and (I.2) we get:

$$R(n) = N(n) \int \varphi(s)\rho(s,n) s ds \ . \tag{II.3}$$

We denote the susceptible population at the beginning by $N = N(0)$.

In the next section we will study the dynamics of the spread of the disease and how the effective reproduction number $R(n)$ depends on the number of infected individual.

medRxiv preprint doi: https://doi.org/10.1101/2020.09.06.20189290; this version posted September 10, 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 .

III. DISEASE SPREAD DYNAMICS

The process starts at step $n = 0$ with one infected individual. At each step that someone is infected, the probability that a was infected is proportional to $S(a)$. Thus,

$$\begin{aligned}&\Pr\left[a \text{ is healthy at step } n\right] = \\ &= \left(1 - \frac{S(a)}{\sum_{b:\, b \text{ is healthy at step } n-1} S(b)}\right) \cdot \\ &\cdot \Pr\left[a \text{ is healthy at step } n-1\right] .\end{aligned} \tag{III.1}$$

We therefore get:

$$\begin{aligned}&\log(\Pr\left(a \text{ is healthy at step } n\right)) - \\ &\quad - \log\left(\Pr\left(a \text{ is healthy at step } n-1\right)\right) = \\ &\quad = -\frac{S(a)}{N(n-1)\mathbb{E}_{b\sim\Lambda_{n-1}}\left[S(b)\right]} \\ &\quad - O\left(\left(\frac{S(a)}{\sum_{b|b \text{ is healthy at step } n-1} S(b)}\right)^2\right) \\ &\quad = -\alpha(n)S(a) - O\left(\left(\frac{\max_b S(b)}{N\mathbb{E}_b S(b)}\right)^2\right) ,\end{aligned} \tag{III.2}$$

Where

$$\alpha(n) = \frac{1}{\sum_{b|b \text{ is healthy at step } n-1} S(b)} . \tag{III.3}$$

Since at each step someone is infected, there can be at most N steps:

$$\begin{aligned}&\log(\Pr\left[\text{a is healthy at step } n\right]) = \\ &= -\left(\sum_{\tau\le n}\alpha(\tau)\right) S(a) - O\left(\left(\frac{\max_b S(b)}{N\mathbb{E}_b S(b)}\right)^2 n\right) = \\ &= -\left(\sum_{\tau\le n}\alpha(\tau)\right) S(a) - O\left(\left(\frac{\max_b S(b)}{\mathbb{E}_b S(b)}\right)^2 \frac{n}{N^2}\right) .\end{aligned} \tag{III.4}$$

Therefore, as long as $\frac{\max_b S(b)}{\mathbb{E}_b S(b)} \ll \sqrt{N}$ we have:

$$\begin{aligned}&\log(\Pr\left[a \text{ is healthy at step } n\right]) \approx \\ &\approx -\left(\sum_{\tau\le n}\alpha(\tau)\right) S(a) .\end{aligned} \tag{III.5}$$

Denote $\delta(n) = \sum_{\tau\le n}\alpha(\tau)$, we get a relation between the susceptibility distributions at steps n and zero:

$$\frac{\rho(s,n)}{\rho(s,0)} = \frac{exp(-\delta(n)s)}{\int \rho(\sigma,0)exp(-\delta(n)\sigma)d\sigma} , \tag{III.6}$$

between the susceptible populations:

$$\frac{N(n)}{N(0)} = \left(\int \rho(s,0)exp(-\delta(n)s)ds\right) , \tag{III.7}$$

and between the effective reproduction numbers (II.3):

$$\frac{R(n)}{R(0)} = \frac{\int \varphi(s)\rho(s,0)exp(-\delta(n)s)sds}{\int \varphi(s)\rho(s,0)sds} . \tag{III.8}$$

This proves Claim I with (I.3) being $1 - \frac{N(n)}{N(0)}$.

We can derive the condition for reaching herd immunity at step n_{herd} by:

$$\begin{aligned}1 = R(n_{herd}) &= N(n_{herd}) \int \varphi(\sigma)\rho(\sigma, n_{end})\sigma d\sigma = \\ &= N(n_{herd}) \int \varphi(\sigma)\rho(\sigma,0)\exp(-\delta(n_{herd})\sigma)\sigma d\sigma ,\end{aligned} \tag{III.9}$$

where $N(n_{herd})$ is related to $N(n=0)$ by (III.7).

IV. HERD IMMUNITY

Consider first the particular case where the infectiousness and susceptibility $I \sim S$ and both are drawn from a Gamma distribution:

$$\rho_{k,\theta}(s,0) = \frac{1}{\Gamma(k)\theta^k} s^{k-1}\exp\left(-\frac{s}{\theta}\right) , \tag{IV.1}$$

where k and θ are shape and scale parameters of the distribution at $n = 0$. Using (III.6) we get:

$$\rho_{k,\theta}(s,n) = \rho_{k,\theta(n)}(s,0), \qquad \theta(n) = \frac{\theta}{1+\theta\delta(n)} . \tag{IV.2}$$

Thus, while the shape of the distribution does not change during the evolution of the disease its scale does.

We denote $\beta(n) = \frac{\theta(n)}{\theta}$, then using (III.7) and (III.8) we get:

$$\frac{N(n)}{N(0)} = \beta(n)^k, \qquad \frac{R(n)}{R(0)} = \left(\frac{N(n)}{N(0)}\right)^{\frac{k+2}{k}} , \tag{IV.3}$$

where

$$R(n) = N(n)\int \rho(s,n)s^2 ds = \left(k^2+k\right)\theta^2(n)N(n) . \tag{IV.4}$$

Herd immunity is reached when $R(n)$ drops below one and this happens when

$$\frac{N(n)}{N(0)} = R^{\frac{-k}{k+2}} , \tag{IV.5}$$

thus proving claim II.

Calculating the fraction of the population ε leading to herd immunity, following claim I, can be carried out analytically for the case of a Gamma distribution, while

medRxiv preprint doi: https://doi.org/10.1101/2020.09.06.20189290; this version posted September 10, 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 .

for general distributions it has do be performed numerically. In figures 1 and 2 we plot ε as a function of the coefficient of variation, i.e. the ratio of the standard deviation and the mean, $\mathrm{Cv} = \frac{\sigma}{\mu}$. We plot the results for three distributions: Gamma, Folded Normal (Truncated Gaussian) and Power Law when R equals three and six, respectively.

We set $\varphi(s) = s$, hence R is the second moment of the distribution. We can see that as Cv approaches 0, the distributions behave similarly. However, for larger values, the behaviour of the system depends upon the distribution with the power law distribution approaching $\varepsilon = 0$ much faster. The coefficient of variation of the truncated Gaussian distribution converges to $\sqrt{\frac{\pi-2}{2}} \approx 0.755$, that is its value for the Half-Normal distribution to which the distribution converges as $\frac{\mu}{\sigma} \to 0^+$.

In general, we see that the higher the variance of the infectiousness and susceptibility the lower the fraction of the population that needs to be infected in order to reach herd immunity.

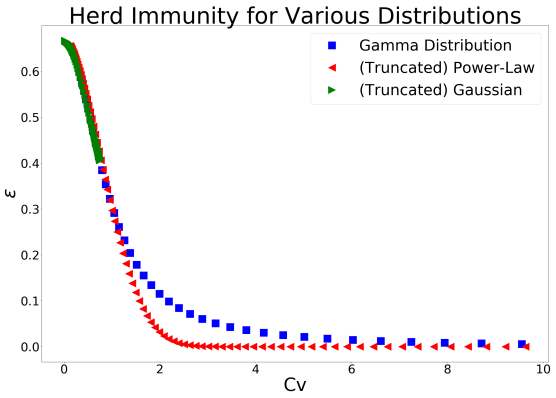

FIG. 1: The percentage of infected in the population necessary for herd immunity to occur as a function of the coefficient of variation for $R = 3$. The results are shown for three distributions: Gamma, Folded Normal and Power Law. The higher the variance of the infectiousness and susceptibility the lower the fraction of the population necessary.

V. DISCUSSION

The spread of the COVID-19 pandemics is characterized by high variance of the infection and susceptible distributions. In addition to the high degree of heterogeneity in infectiousness and susceptibility, one expects a significant correlation between them stemming, for instance, from the social aspect of the spread of diseases. We studied the implications of this structure on the condition to reach herd immunity.

We proved two claims, one for general distributions and one for the Gamma distribution and showed that the heterogeneity and correlation have a drastic effect on the estimate of the percentage of the population that must contract the disease before herd immunity is reached.

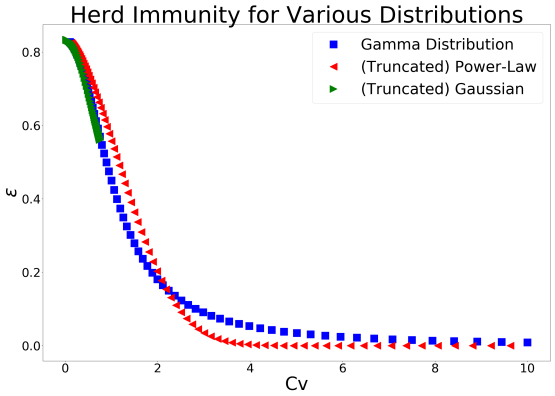

FIG. 2: The percentage of infected in the population necessary for herd immunity to occur as a function of the coefficient of variation for $R = 6$. The results are shown for three distributions: Gamma, Folded Normal and Power Law. In comparison to figure 1 we see, as expected, that the required fraction is higher for a given coefficient of variation.

Under the assumption of Gamma distribution we found that for COVID-19 a fraction $\varepsilon \approx 5\%$ of infected population suffices to reach herd immunity while a fraction $\varepsilon \approx 9\%$ is needed for COVID-2. While writing the paper we became aware of two recent works [9, 10] that, using different mathematical fraemeworks, reached similar conclusions for the fraction of infected population that is required for herd immunity under the assumption of Gamma distribution.

Our mathematical analysis of the disease spread dynamics can be viewed as a random walk on a complete graph. It is of interest to study other graph structures and quantify the differences [11].

Acknowledgements

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

[1] K Rock, S Brand, J Moir, MJ Keeling, "Dynamics of infectious diseases", Reports on Prog. Phys. **77**, 026602 (2014).
[2] R. Pastor-Satorras, C. Castellano, P. Van Mieghem and

medRxiv preprint doi: https://doi.org/10.1101/2020.09.06.20189290; this version posted September 10, 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. 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", https://doi.org/10.1101/2020.05.21.20104521.

[5] A Endo, S Abbott, A Kucharski, S Funk, Estimating the overdispersion in covid-19 transmis- sion using outbreak sizes outside china. Wellcome Open Res. **5**, 67 (2020).

[6] Y. Oz, I. Rubinstein and M. Safra, "Superspreaders and High Variance Infectious Diseases", arXiv:2008.07352 [q-bio.PE].

[7] 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).

[8] A. V. Tkachenko, S. Maslov, A. Elbanna, George. N. Wong, Z. J. Weiner, N. Goldenfeld, "Persistent heterogeneity not short-term overdispersion determines herd immunity to COVID-19", https://doi.org/10.1101/2020.07.26.20162420.

[9] J. Lourenco, F. Pinotti, C. Thompson, S. Gupta, "The impact of host resistance on cumulative mortality and the threshold of herd immunity for SARS-CoV-2", https://doi.org/10.1101/2020.07.15.20154294.

[10] R Aguas, et al., Herd immunity thresholds for sars-cov-2 estimated from unfolding epidemics. medRxiv 2020.07.23.20160762 (2020).

[11] Y. Oz, I. Rubinstein and M. Safra, work in progress.