The SIR Epidemic Model

Charles S. Peskin
Courant Institute of Mathematical Sciences, New York University
May 9, 2020

This is an introduction to the SIR epidemic model. Our purpose is not to assess the applicability of the model to the real world, although we do want to make the underlying assumptions of the model clear, but rather to describe the model's interesting mathematical behavior and real-world implications, insofar as it may be applicable. The model is defined by the following differential equations:

$$\frac{dS}{dt} = -aS\left(\frac{I}{N}\right), \tag{1}$$

$$\frac{dI}{dt} = +aS\left(\frac{I}{N}\right) - bI, \tag{2}$$

$$\frac{dR}{dt} = +bI. \tag{3}$$

Here S is the number of susceptible people, I is the number of infected (and therefore infectious) people, and R is the number of recovered (and therefore immune) people. The sum

$$N = S + I + R \tag{4}$$

is the total number of people, and this is a conserved quantity, since the model does not consider births or deaths.

The parameters of the model, a and b, have units of 1/time. In a short time interval Δt, $a\Delta t(I/N)$ is the probability that a given susceptible person becomes infected, and $b\Delta t$ is the probability that a given infected person recovers. Strictly speaking, these statements are only true in the limit $\Delta t \to 0$, and then the probabilities also become zero, so what we really mean is that

$$a\left(\frac{I}{N}\right) = \lim_{\Delta t \to 0} \frac{P(\text{infection})}{\Delta t}, \tag{5}$$

$$b = \lim_{\Delta t \to 0} \frac{P(\text{recovery})}{\Delta t}. \tag{6}$$

where $P(\text{infection})$ is the probability that any *one* susceptible person becomes infected during a time interval of duration Δt, and $P(\text{recovery})$ is the probability

that any *one* infected person recovers during a time interval of duration Δt. A shorter way to say this is that $a(I/N)$ is the *probability per unit time* that a given susceptible person becomes infected, and b is the *probability per unit time* that a given infected person recovers. Note the unrealistic assumption here that the probability per unit time of recovery is constant, independent of how long the infected person has been infected.

The infectivity parameter a can be factored as follows:

$$a = a_0 \, p_{\text{trans}}, \tag{7}$$

where a_0 is the number of encounters per unit time that any one susceptible person has, where an *encounter* is defined as an interaction with another person that could result in the susceptible person's becoming infected, and where p_{trans} is the probability that infection of the susceptible person will actually happen *if* the other participant in the encounter is infected. Note that the probability of this last condition being relized (in a homogenous well-mixed population, with no influence of infection on behavior) is I/N, and this explains why a has to be multiplied by I/N to get the probability per unit time that any one susceptible person will become infected.

As a practical matter, when setting parameters, it is best to set b first by keeping in mind that $1/b$ is the mean time from infection to recovery, and then to set a by using the fact that a/b is the *reproduction number* of the disease. The reproduction number is the average number of infections that will be caused *directly* by transmission from a single infected person in an otherwise susceptible population. Note that the reproduction number does not count infections caused indirectly by the first infected person via some other person, only the infections caused by encounters involving that particular first infected person. The reason that a/b is the reproduction number is that by definition the rest of the population is susceptible, so $S = N$ (we ignore here the distinction between N and $N-1$), and since we are only counting infections caused by one person, $I = 1$. Thus, the rate at which the one infected person is causing new infections is $aN(1/N) = a$, and the amount of time that this one infected person has in which to do this is, on average, $1/b$, so a/b is the average number of people directly infected.

Although we have defined the reproduction number by considering what happens with one infected person in an otherwise susceptible population, it is also important to recognize that the reproduction number is effectively reduced when some fraction of the population is recovered and therefore immune. This is the phenomenon of *herd immunity*. The same argument as above shows that the number of cases caused directly by one infected person in a population containing a

mixture of susceptible and recovered people is $aS(1/N)/b$, which is a the original reproduction number multiplied by the fraction S/N. If $(a/b)(S/N) < 1$, no epidemic will occur. This result means that it is possible to protect an *entire* population (in a sense that will be made precise later) by vaccinating only a *fraction* of the population. The benefit of herd immunity, however, is limited when the original reproduction number is too large. If $a/b = 2$, then we can prevent an epidemic by vaccinating only slightly more than half of the population, but if $a/b = 10$ we have to vaccinate more than 90% of the population to prevent an epidemic.

The behavior of the solutions of the differential equations of the SIR model is very different depending on whether $a/b < 1$ or $a/b > 1$. Of course there is a borderline case $a = b$, and as a practical matter the behaviors blurr into each other when a and b are approximately equal, but we will not discuss the borderline situation here. We call $a/b < 1$ the *non-epidemic case* and $a/b > 1$ the *epidemic case*. This terminology refers to the situation in which no one is intially immune. When a fraction of the population is initially immune, and therefore when the susceptible fraction is less than one, there is still a non-epidemic case and an epidemic case but the threshold is different — as explained above the relevant quantity is then the effective reproduction number $(a/b)(S(0)/N)$. In the following two sections, however, we consider the case in which no one is initially immune.

Non-Epidemic Case

Here $a < b$, and we assume that some number $I(0)$ of infected people are introduced into an otherwise susceptible population, so $R(0) = 0$. We assume that $I(0) << N$. Thus, for practical purposes $S(0) = N$, and we make the provisional assumption that this approximation remains valid for all time, i.e., that $I(t) + R(t) << N$, for all t. The self-consistency of this assumption will be checked later.

With $S = N$, the equation for $I(t)$ becomes

$$\frac{dI}{dt} = (a - b)I = -(b - a)I, \tag{8}$$

and of course this has the solution

$$I(t) = I(0)\exp(-(b - a)t) \tag{9}$$

The quantity $R(\infty)$ measures the total number of people who ever were infected (including those who were initially infected). Since $R(0) = 0$, $R(\infty)$ can be evaluated as

$$R(\infty) = \int_0^\infty bI(t)dt = I(0)\frac{b}{b-a} = \frac{I(0)}{1-(a/b)}. \tag{10}$$

Our approximation is valid as long as $I(0)/(1-(a/b)) << N$, and of course a/b has to be in $(0,1)$, since we are here concerned with the non-epidemic case. Note that $R(\infty)$ blows up as $(a/b) \to 1$, which is an indication that the non-epidemic case is blurring into the epidemic case, which will be discussed below.

In the non-epidemic case, the total number of people who were ever infected is:

- proportional to $I(0)$
- independent of N

This means that the non-epidemic is a local phenomenon. Even if $I(0)$ is multiplied by a large number (for example, if a/b is 0.9, $R(\infty) = 10\,I(0)$), the non-epidemic event does not have any noticeable impact on the population as a whole. This is what was meant when we said that herd immunity can protect the whole population even though only a part of the population is immnune. The point is that herd immunity can ensure that any clusters of cases that occur involve numbers that do *not* scale up with the population size N.

Epidemic Case

Here $a > b$, but otherwise we consider the same situation as before, in which some number $I(0)$ of infected people are introduced into an otherwise susceptible population (so $R(0) = 0$), with $I(0) << N$. At early times, when S is approximately equal to N we have the same dynamics as before:

$$I(t) = I(0)\exp\left((a-b)t\right), \tag{11}$$

but now, since $a > b$, this is exponential growth instead of exponential decay. Clearly, such exponential growth cannot continue indefinitely, since it would soon predict that there are more infected people than people in the population, and well before that, the approximation that S is approximately equal to N would become invalid.

To see what actually happens, we can of course solve the system (1-3) numerically, but we can also obtain some analytic results in the following way.

First, we rewrite the basic differential equations of the model in terms of fractions of the population. Let

$$\sigma = S/N, \;\; \iota = I/N, \;\; \rho = R/N. \tag{12}$$

Then, dividing both sides of each of the equations (1-3) by N and recalling that N is constant, we get

$$\begin{aligned} \frac{d\sigma}{dt} &= -a\sigma\iota, & (13)\\ \frac{d\iota}{dt} &= +a\sigma\iota - b\iota, & (14)\\ \frac{d\rho}{dt} &= +b\iota. & (15)\\ & & (16) \end{aligned}$$

Note that the parameter N has conveniently dissappeared. This is because we used the formulation of the SIR model in which the infection rate depends on I through I/N. The lack of dependence on N in the above equations implies that our model is scale-invariant. When conclusions are expressed in terms of fractions of the population, they will be the same no matter how large the population may be.

A useful trick for the analysis of the above system is to eliminate time and use ρ as the independent variable. This is achieved by dividing $d\sigma/dt$ by $d\rho/dt$. The result, after canceling ι, is

$$\frac{d\sigma}{d\rho} = -\left(\frac{a}{b}\right)\sigma, \tag{17}$$

which is a differential equation that we can easily solve for σ as a function of ρ. Since we are assuming that the epidemic starts with no one immune, ρ starts at zero and increases from there. The value of σ at the start of the epidemic is *very* close to 1. For example, in a city of a million people, there may be one or two people infected, so σ has a value like 0.999998, for example, and we may as well set $\sigma(0) = 1$. With this initial condition, we have

$$\sigma(\rho) = \exp\left(-\left(\frac{a}{b}\right)\rho\right). \tag{18}$$

Several important characteristics of the epidemic can be deduced from this relationship, as we discuss in the following.

During an epidemic, the fraction of the population that is infected rises to a maximum and then declines, and we would like to know what the maximum infected fraction will be. To evaluate this, we note that

$$\sigma + \iota + \rho = 1, \tag{19}$$

so maximizing ι is the same as minimizing $\sigma + \rho$. Accordingly, we set

$$\begin{aligned} 0 &= \frac{d}{d\rho}\left(\sigma(\rho) + \rho\right), && (20)\\ &= -\left(\frac{a}{b}\right)\sigma(\rho) + 1, && (21)\\ &= -\left(\frac{a}{b}\right)\exp\left(-\left(\frac{a}{b}\right)\rho\right) + 1. && (22)\\ & && (23) \end{aligned}$$

The second and third lines of the above equation are both useful. Fron the second line, we get the value of σ when the infected fraction is at its peak, and from the third line we get the corresponding value of ρ. These values are

$$\begin{aligned} \sigma &= \frac{1}{(a/b)}, && (24)\\ \rho &= \frac{\log(a/b)}{(a/b)}, && (25) \end{aligned}$$

and therefore, the fraction of the population that is infected at any one time has the maximum value

$$\iota_{\max} = 1 - \frac{1 + \log(a/b)}{(a/b)}. \tag{26}$$

Note that the maximum infected fraction depends only on the reproduction number (a/b). It is an increasing function of (a/b), with the value 0 when $(a/b) = 1$, and approaching 1 as $(a/b) \to \infty$.

In a similar way, one could evaluate the maximum of the rate at which new infections are occurring. The fraction of the population that is newly infected per unit time (that is, the number of new infections per unit time divided by N) is given by $a\sigma\iota = a\sigma(1 - (\sigma + \rho))$, and the maximum value of this expression can be found by making use of equation (18) in much the same way as in the foregoing calculation of $\iota_{\max}$, but we leave the details as an exercise for the reader.

Another quantity of interest is the fraction of the population that is recovered and therefore immune at the end of the epidemic. This is the same as the fraction

of the population consisting of people who became infected at any time during the epidemic. Thus it measures the total impact of the epidemic on the population, and it is also important for the future herd immunity of the population, in case the population is challenged by the same pathogen again. The recovered fraction at the end of the epidemic can be obtained from equation (18) by noting that $\iota = 0$ at the end of the epidemic, so $\sigma + \rho = 1$, or $\sigma = 1 - \rho$. This gives the equation

$$\exp\left(-\left(\frac{a}{b}\right)\rho\right) = 1 - \rho. \tag{27}$$

One solution of this equation is $\rho = 0$, and that is the only non-negative solution if $a < b$, which shows again that a non-epidemic has a negligeable effect on the population as whole. If $a > b$, however, it is easy to see by plotting the the two sides of (27) as functions of ρ that there is another solution which has $0 < \rho < 1$, and this solution corresponds to the recovered fraction at the end of the epidemic. Like the maximum infected fraction, it depends only on the reproduction number a/b. Even though we do not have a formula for this dependence, it is easy to determine it by solving equation (27) numerically. The solution of (27) for the recovered fraction of the population at the end of the epidemic is plotted in Figure 1 as a function of the reproduction number, along with the fraction of the population that is infected at the peak of the epidemic, as given by equation (26).

The fact that we get a definite results for the epidemic case without having to specify the number of people initially infected shows that in the epidemic case the number of initially infected people does not mattter. The initial infections are the spark that ignites the fire, and if the fire is indeed ignited it will run its course independent of the intensity of the spark. The only difference that the number of people initially infected will make is a shift on the time axis, but the fraction of the population that is infected at the height of the epidemic, and the fraction of the population that is ever infected during the course of the epidemic, will not depend at all on the number of initial infections that set the epidemic going (assuming that number was too small to be itself a significant fraction of the whole population). Also, since the results we have calculated are fractions of the population, the corresponding numbers of people scale with the population size. Thus, in the epidemic case, the numbers of people affected are

- independent of $I(0)$
- proportional to N

and this is exactly opposite to the situation in the non-epidemic case!

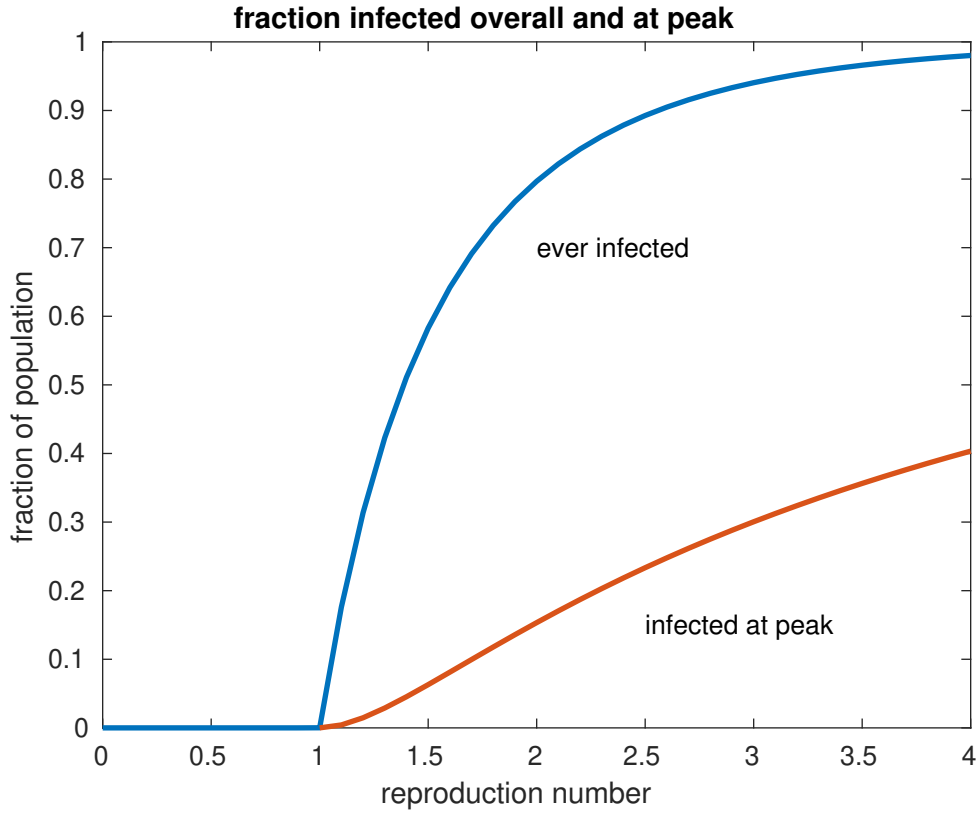

Figure 1: Infected fraction of the population as a function of the reproduction number. Upper curve shows the fraction of the population consisting of people who were infected at any time during the course of the epidemic, and lower curve shows the fraction of the population that is infected at the peak of the epidemic, i.e., at the time at which the number of currently infected people is the largest.

When Part of the Population is Initially Immune: Non-Epidemic Case

Here we assume that a small number $I(0)$ of infected people are introduced into a large population consisting of $S(0)$ susceptible people and $R(0)$ recovered (and therefore immune) people. Since $I(0)$ is a small number, and $S(0)$ and $R(0)$ are large, we may write

$$N = S(0) + R(0), \tag{28}$$

and we make the provisional assumption that $S(t)$ and $R(t)$ stay close to their initial values for all t. In that case, we have the equation

$$\frac{dI}{dt} = a\left(\frac{S(0)}{N}\right) I - bI, \tag{29}$$

and this can be written

$$\frac{dI}{dt} = (a\sigma(0) - b)\, I, \tag{30}$$

where

$$\sigma(0) = \frac{S(0)}{N} = \frac{S(0)}{S(0) + R(0)}. \tag{31}$$

This shows that the infectivity parameter a is effectively reduced by being multiplied by the susceptible fraction of the population $\sigma(0)$. Thus, the condition for the non-epidemic case is now

$$a\sigma(0) < b, \tag{32}$$

and in that case we have

$$I(t) = I(0) \exp\left(-\left(b - a\sigma(0)\right)\right), \tag{33}$$

$$R(\infty) - R(0) = \frac{I(0)}{1 - (a\sigma(0)/b)}. \tag{34}$$

These are exactly the same results as before, except that the reproduction number a/b has been replaced by the smaller effective reproduction number $a\sigma(0)/b$. In particular, the non-epidemic is still a local phenomenon, in the sense that the number of infected people is proportional to the number initially infected and independent of the population size. The benefit of having only part of the population initially susceptible is two-fold. First, and most important, it is now easier for the non-epidemic case to be what actually happens, since the condition $a\sigma(0) < b$ is less restrictive than $a < b$. Also, even if $a < b$ so that we would be in the non-epidemic case without anyone initially immune, the number of infections (including those initially infected) is reduced from $I(0)/(1-a/b)$ to $I(0)/(1-a\sigma(0)/b)$.

When Part of the Population is Initially Immune: Epidemic Case

Here we consider the same situation as in the previous section, but with $a\sigma(0) - b > 0$, so that (33) describes exponential growth. The "0" in $\sigma(0)$ refers to $t = 0$, but here we are going to think of σ as a function of ρ, and the value of ρ at $t = 0$ is not zero, so from now on we write σ_0 and ρ_0 as the intial values of σ and ρ, and we note that

$$\sigma_0 + \rho_0 = 1, \tag{35}$$

since the number of initially infected people is negligible as a fraction of the population. The solution of (18) for $\sigma(\rho)$ is now

$$\sigma(\rho) = \sigma_0 \exp\left(-\left(\frac{a}{b}\right)(\rho - \rho_0)\right), \tag{36}$$

and the domain of ρ is (ρ_0, ρ_∞), where $\rho_\infty < 1$ remains to be determined. Note that $\rho_\infty - \rho_0$ is the fraction of the population that is ever infected during the course of the epidemic.

In much the same way as before, we can use (36) to determine $\iota_{\max}$ and $\rho_\infty - \rho_0$. To find $\iota_{\max}$, we minimize $\sigma(\rho) + \rho$:

$$\begin{aligned} 0 &= \frac{d}{d\rho}\left(\rho + \sigma_0 \exp\left(-\left(\frac{a}{b}\right)(\rho - \rho_0)\right)\right), && (37)\\ &= 1 - \left(\frac{a}{b}\right)\sigma, && (38)\\ &= 1 - \left(\frac{a}{b}\right)\sigma_0 \exp\left(-\left(\frac{a}{b}\right)(\rho - \rho_0)\right), && (39) \end{aligned}$$

and from this we find that the values of σ and ρ at the peak of the epidemic are

$$\begin{aligned} \sigma &= \frac{1}{(a/b)}, && (40)\\ \rho &= \rho_0 + \frac{\log\left((a/b)\sigma_0\right)}{(a/b)}, && (41) \end{aligned}$$

It follows that

$$\begin{aligned} \iota_{\max} &= 1 - \rho_0 - \frac{1 + \log\left((a/b)\sigma_0\right)}{(a/b)}, && (42)\\ &= \sigma_0\left(1 - \frac{1 + \log\left((a/b)\sigma_0\right)}{\left((a/b)\sigma_0\right)}\right). && (43) \end{aligned}$$

When the entire population was initially susceptible, we had the special case of this formula with $\sigma_0 = 1$. In comparison to that case, we see that $\iota_{\max}$ is here reduced in two ways. First, and most important, the reproduction number a/b is here replaced by the effective reproduction number $(a/b)\sigma_0$, which is smaller. The second effect is simply the leading scale factor σ_0. In summary, at the peak of the epidemic, the infected population will be a certain fraction of the initially susceptible population, not of the whole population, and moreover that infected fraction of the initially susceptible population will be as if the reproduction number had been reduced by being multiplied by the initially susceptible fraction of the population.

Now we consider the fraction of the population that is ever infected during an epidemic in which some fraction of the population is initially immune. At the end of the epidemic, $\iota = 0$, so $\sigma + \rho = 1$. Combining this with (36), we get

$$1 - \rho_\infty = \sigma_0 \exp\left(-\left(\frac{a}{b}\right)(\rho_\infty - \rho_0)\right), \tag{44}$$

and we also have

$$1 - \rho_0 = \sigma_0. \tag{45}$$

Combining these equations gives the result

$$\rho_\infty - \rho_0 = \sigma_0 \left(1 - \exp\left(-\left(\frac{a}{b}\right)(\rho_\infty - \rho_0)\right)\right), \tag{46}$$

which can also be written as

$$\left(\frac{\rho_\infty - \rho_0}{\sigma_0}\right) = 1 - \exp\left(\left(-\sigma_0 \frac{a}{b}\right)\left(\frac{\rho_\infty - \rho_0}{\sigma_0}\right)\right). \tag{47}$$

Now $\rho_\infty - \rho_0$ is the fraction of the population that ever becomes infected (without regard to when) over the course of the epidemic, and of course all of these people were initially susceptible. The quantity $(\rho_\infty - \rho_0)/\sigma_0$ is the fraction of the initally susceptible population that becomes infected over the course of the epidemic. This fraction obeys the same equation as was found previously when the whole population was initially susceptible, except that here the reproduction number a/b is replaced by the effective reproduction number $(a/b)\sigma_0$. As in the case of $\iota_{\max}$, there is a two-fold benefit to having part of the population be initially immune. First, there is effectively a reduction in the reproduction number, and second, the relevant population (i.e., those initially susceptible) is smaller, so the number of cases is fewer. This two-fold benefit is illustrated in Figure 2 for the particular case in which $1/3$ of the population is initially immmune.

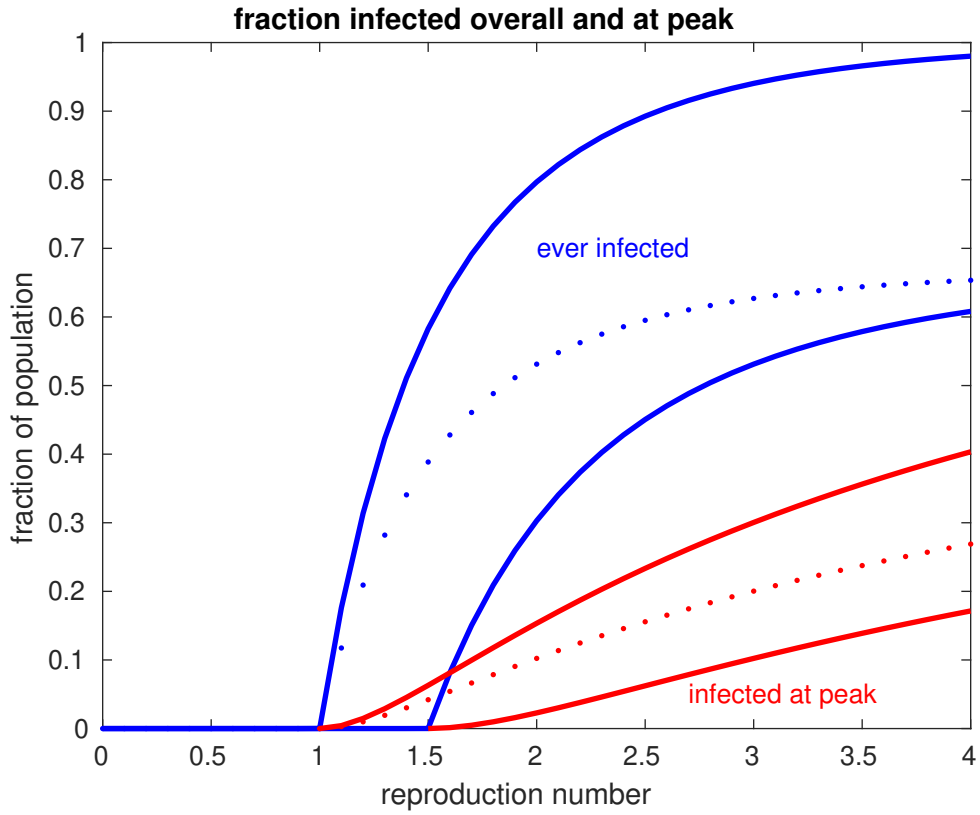

Figure 2: The benefit of having 1/3 of the population initially immune. Blue curves show fraction of population ever infected over the course of the epidemic, and red curves show the fraction infected at the peak of the epidemic, as functions of the reproduction number a/b. In both cases the upper solid curve is what happens when no one is initially immune (same as in Figure 1), and the lower solid curve is what happens when 1/3 of the population is initially immune. The dotted curves are simply 2/3 of the upper solid curves, so they show the direct benefit of having only 2/3 of the population initially susceptible. The difference between each dotted curve and the corresponding lower curve is the indirect benefit, i.e., the herd immunity effect.