6 Population Structure and Contacts

Thus far, we have assumed that the population is divided into compartments of S, I, or R, and, perhaps, their vaccinated equivalents S_v and I_v. Throughout, we have assumed that the contact rate c is the same for everyone. This means that each person in our model comes into contact with the same number of people every unit of time. In turn, the compartment affiliation of those contacts is in proportion to the compartment's share of the total population. For example, in a scenario where $c = 2$ and the population is split 60%/40% between S and I, respectively, we assume that any person in any compartment comes into contact with 2 people per time, each of whom is susceptible with probability 0.6 and infected with probability 0.4.

This week we'll ask: what if there are patterns in the population's contact rates? For example, suppose there are adults and children, and each child contacts 20 other children and 10 adults per day, while each adult contacts 10 other adults and 2 children per day. How would a disease spread in this sort of structured population? What do we need to do to modify our previous modeling approach? And, with modifications implemented, what questions can we ask and answer that we previously could not?

6.1 The Contact Matrix

Suppose that our population can be divided into two demographic groups, indexed 1 and 2. We'll develop a method that is generic to the labels we give to the groups, but keep a few examples of population pairs in the back of your mind, like children and adults, or essential workers and non-essential workers.

Now consider that a person in S_1, the group 1 susceptibles, can become infected due to contact with someone from either I_1 or I_2. Thus, while we previously wrote

$$\dot{S}(t) = -\left(c\frac{pI}{N}\right)S$$

for a single population, we'll now write

$$\dot{S}_1(t) = -\left(c_{11}\frac{pI_1}{N_1}\right)S_1 + -\left(c_{12}\frac{pI_2}{N_2}\right)S_1\ ,$$

where c_{11} is the rate at which people in group 1 experience contacts from others in group 1, and c_{12} is the rate at which people in group 1 experience contacts from those in group 2. We can write this more compactly as

$$\dot{S}_1(t) = -\left(c_{11}\frac{I_1}{N_1} + c_{12}\frac{I_2}{N_2}\right)pS_1$$

and then, generically for any group m as

$$\dot{S}_m(t) = -pS_m\sum_{j=1,2}c_{mn}\frac{I_n}{N_n}\ . \tag{1}$$

Note that we can also obtain these equations in population proportions easily as well by dividing both sides by N to get

$$\dot{s}_m(t) = -p s_m \sum_{j=1,2} c_{mn} \frac{i_n}{\omega_n} \,. \tag{2}$$

where $\omega_n = N_n/N$ is the proportion of the total population that is in group n.[1]

Finally, we might also observe that the equation above has the form of matrix vector multiplication,[2] and that we can also write out the other equations in matrix-vector form, too.

$$\begin{aligned}
\dot{\mathbf{s}} &= -p \left(D_{\mathbf{s}} C D_{\omega}^{-1}\right) \mathbf{i} \\
\dot{\mathbf{i}} &= p \left(D_{\mathbf{s}} C D_{\omega}^{-1}\right) \mathbf{i} - \gamma \mathbf{i} \\
\dot{\mathbf{r}} &= \gamma \mathbf{i}
\end{aligned} \tag{3}$$

where $D_{\mathbf{s}}$ represents a diagonal matrix with the entries of $\mathbf{s}$ on the diagonal, and D_{ω}^{-1} represents a diagonal matrix with entries of ω_m^{-1} on the diagonal.[3]

These equations are remarkable because they helps us to see clearly that the matrix C is the important novel piece. This general form for the equations is also appropriate regardless of whether we have two groups or we have twenty. You might also notice that if there is only one group, then the $\mathbf{i} \to i$, $D_{\mathbf{s}} \to s$, and $D_{\omega}^{-1} \to 1$, in which case we recover the original SIR equations written in Lecture 1.

Let's now explore how the contact matrix can affect dynamics via simulation. Here we'll consider $p = 1$ and $\gamma = 1$ with two populations of 1000 each, beginning with one infected person in each population. We'll consider three contact matrices, results for which are displayed left to right in Figure 1.

$$C = \begin{pmatrix} \mathbf{1.5} & 0 \\ 0 & \mathbf{1} \end{pmatrix} \qquad C = \begin{pmatrix} \mathbf{1.5} & 0.1 \\ 0.1 & \mathbf{1} \end{pmatrix} \qquad C = \begin{pmatrix} \mathbf{1.5} & 0.1 \\ 0.1 & \mathbf{0.5} \end{pmatrix} \,.$$

Let's interpret these matrices. The boldfaced numbers on the diagonals are C_{ii} for $i = 1, 2$, meaning that they tell us how a population contacts itself. In other words, the diagonal entries tell us about within-group contact rates. That means that the off-diagonal entries tell us about between group contact rates. Note then that the off-diagonal entries of the left matrix are zeroes. This means that the left matrix reflects zero contact between the two groups, and we can therefore think of this scenario as two distinct populations that never mix. As a result, Figure 1(left) shows dynamics for the two populations that are identical to what we'd get if we used a single-population model and simply simulated twice. In particular, $R_0 = 1.5$ for group one,

[1] Try dividing both sides of Eq. (1) by N, and you'll see that we need the ω terms to make things work.

[2] Recall that $x_m = \sum_n A_{mn} y_n$ is the per-term version of the matrix-vector notation of $\mathbf{x} = A\mathbf{y}$.

[3] Why can we use this notation? Interestingly, to compute the inverse of a diagonal matrix, one can simply invert the entries on the diagonal. Can you show why this is so?

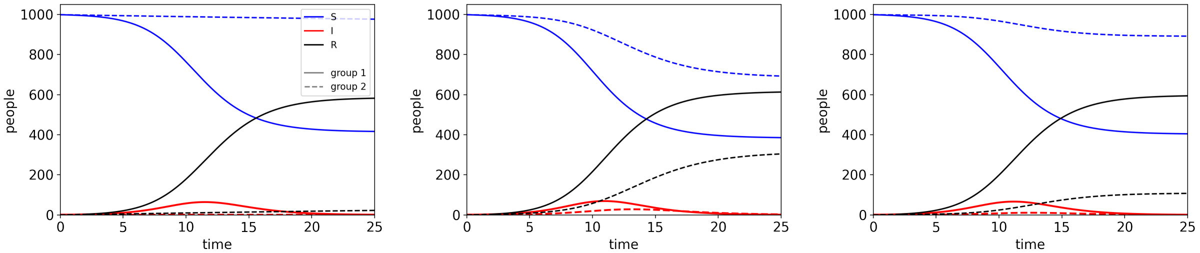

Figure 1: SIR simulations are shown for varying contact matrices C, but otherwise identical parameters and initial conditions. Note that the left plot corresponds to two populations in isolation from each other. Leaving all else constant, the middle plot then introduces a small amount of contact between the two populations. Then, leaving all else constant, the right plot decreases the rate of contact only within group two. Notice that even though population two is incapable of sustaining an epidemic *on its own*, it experiences an epidemic because of its coupling with group two.

and $R_0 = 1$ for group two. Notice that as expected, there is a large epidemic in group one but not group two.

To the first scenario, the middle matrix adds a low level of between-group contacts. As a result, the outcomes change substantially. Not only do both populations experience epidemics, but the epidemic in group 1 is even larger. All of this is a consequence of the fact that now, the two groups in the population are coupled to each other.

To the second scenario, the middle matrix decreases the level of within-group contact in group two from 1 to 0.5. This means that in the absence of any connection with group one, the dynamics in group two would be characterized by $R_0 = 0.5$, a number far less than one. Nevertheless, 107 people (10.7%) in group two are infected by $t = 25$. in fact, in terms of our equations developed in Lecture 2 for the final epidemic size in a single population, this would be as if group two experienced an epidemic with $R_0 \approx 1.06$.

6.2 Aside: discrete-time linear dynamics

I want you to put infectious diseases out of your head for a moment, and consider a population of two kinds of cells, y and z. In each time step, each y creates a copy of itself and a copy of z. And, each z creates one copy of itself, and two copies of y. After creating these copies, the originals die. What is the long-term growth rate of the overall population?

To start, let's write down our dynamical equations:

$$\begin{aligned} y(t+1) &= y(t) + 2z(t) \\ z(t+1) &= y(t) + z(t) \end{aligned} \tag{4}$$

We could also simply write

$$\begin{pmatrix} y(t+1) \\ z(t+1) \end{pmatrix} = \begin{pmatrix} 1 & 2 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} y(t) \\ z(t) \end{pmatrix} \qquad \text{or} \qquad \mathbf{x}(t+1) = A\mathbf{x}(t), \quad A = \begin{pmatrix} 1 & 2 \\ 1 & 1 \end{pmatrix} \tag{5}$$

The last equation is helpful, because it tells us that to update our population counts held in the vector $\mathbf{x}$, we need only multiply by the matrix A once more. This matrix A therefore updates our population counts from one generation to the next. We might even call it a next-generation matrix. This allows us to write out our dynamics from initial conditions

$$\mathbf{x}(t) = A^t\mathbf{x}(0)\ . \tag{6}$$

How much does the population grows from one generation to the next? There are many ways to do this computation, but one of my favorites is to recall that, for a particular set of conditions on our matrix A, the eigenvectors of A form a basis.[4] This means that we could take our initial conditions and write them as a mixture of A's eigenvectors,

$$\begin{pmatrix} y(0) \\ z(0) \end{pmatrix} = \alpha_1\mathbf{u}_1 + \alpha_2\mathbf{u}_2$$

From here, we can multiply by A to update our population sizes from $t = 0$ to $t = 1$ as follows

$$\begin{aligned} \mathbf{x}(1) &= A\mathbf{x}(0) \\ &= A\big(\alpha_1\mathbf{u}_1 + \alpha_2\mathbf{u}_2\big) \\ &= \alpha_1 A\mathbf{u}_1 + \alpha_2 A\mathbf{u}_2 \\ &= \alpha_1\lambda_1\mathbf{u}_1 + \alpha_2\lambda_2\mathbf{u}_2 \end{aligned} \tag{7}$$

If we multiplied again by A we could update from $t = 1$ to $t = 2$,

$$\begin{aligned} \mathbf{x}(2) &= A\mathbf{x}(1) \\ &= A\big(\alpha_1\lambda_1\mathbf{u}_1 + \alpha_2\lambda_2\mathbf{u}_2\big) \\ &= \alpha_1\lambda_1 A\mathbf{u}_1 + \alpha_2\lambda_2 A\mathbf{u}_2 \\ &= \alpha_1\lambda_1^2\mathbf{u}_1 + \alpha_2\lambda_2^2\mathbf{u}_2 \end{aligned} \tag{8}$$

In fact, we can now just jump right to the general expression, because we can see that each timestep means multiplying by the matrix A, which simply increases the degree on the eigenvalues λ_1 and λ_2,

$$\mathbf{x}(t) = \alpha_1\lambda_1^t\mathbf{u}_1 + \alpha_2\lambda_2^t\mathbf{u}_2 \tag{9}$$

[4] When a set of vectors forms a basis for our space, that means that you can express an arbitrary vector as a linear combination of the basis vectors.

Now, suppose that $\lambda_1 > \lambda_2$, i.e. that λ_1 is the largest eigenvalue of A. Then,

$$\mathbf{x}(t) = \lambda_1^t \left(\alpha_1 \mathbf{u}_1 + \alpha_2 \left[\frac{\lambda_2}{\lambda_1} \right]^2 \mathbf{u}_2 \right) \; . \tag{10}$$

As t grows large, the term $[\lambda_1/\lambda_2]^t$ shrinks to zero, and we therefore have

$$\mathbf{x}(t) \sim \lambda_1^t \alpha_1 \mathbf{u}_1 \; , \tag{11}$$

an equation which tells us that the long-term growth factor of the population is λ_1 per iteration, *and* the long-term population is proportional to the entries of $\mathbf{u}_1$, the eigenvector corresponding to λ_1.

Before we move on, I'd like to reflect on what we have just shown. We have shown that if one has a linear dynamics of the form $\mathbf{x}(t+1) = A\mathbf{x}(t)$, then the growth rate corresponds to the largest eigenvalue of A. What we will do next is to consider a similar system: instead of two types of cells, we'll consider an arbitrary number of infectious groups. Instead of writing down how the cells in each group create more cells of the same and different groups, we'll consider how infections in each group create more infections in the same and different groups. In other words, we're going to write down a matrix-type equation that tracks infections from generation to generation, and when we do, we're going to inspect its largest eigenvalue!

6.3 R_0 with population structure: the next-generation matrix

How can we calculate R_0 in a population where there is non-trivial contact structure? In fact, what if we had, say, twenty groups in the population, and thus we had $20 \times 20 = 400$ different values in the matrix C? How in the world would we be able to compute the rate at which infections grow in an otherwise susceptible population? To answer these questions, we will now develop the **next generation matrix**, abbreviated NGM, and for which we'll use the letter G.

The next generation matrix, like R_0, takes us from the realm of rates (measured per time) into generations of infection (measured per infection). The idea behind the NGM is to ask: *how do infection counts in* **this** *generation lead to infection counts in the* **next** *generation?* To compute this, we're going to consider two different processes, called transmission and transition. Transmission accounts for the creation of new infections. Transition accounts for the movement of existing infections from one group to another. It may be helpful for you to think of these processes as tracking the demographics of the infections themselves: Transmission tracks the births of new infections, while Transition tracks their aging and eventual death.

In thinking about transmission and transition, we're going to consider the **infection subsystem** of our larger system. The infection subsystem refers to only those compartments that track those who are newly infected or are already infected and remain so. In the case of our simply two-group example, the infected subsystem would consist of compartments I_1 and I_2. Let's write out their equations to have them on hand,

$$\begin{aligned}\dot{I}_1 &= pS_1c_{11}\frac{I_1}{N_1} + pS_1c_{12}\frac{I_2}{N_2} - \gamma I_1\\ \dot{I}_2 &= pS_2c_{21}\frac{I_1}{N_1} + pS_2c_{22}\frac{I_2}{N_2} - \gamma I_2\ . \end{aligned} \tag{12}$$

To derive the NGM, we're going to write down the Jacobian of the infected subsystem, and decompose it. A **matrix decomposition** is a way of saying that we're going to take one matrix and rewrite it as the sum or product of two or more other matrices. Here, the two elements of our additive decomposition will be T (for transmission) and Σ (for transition). Taking the partial derivatives of Eq. (12), we get the Jacobian

$$J = \begin{pmatrix} p\frac{S_1}{N_1}c_{11} - \gamma & p\frac{S_1}{N_2}c_{12} \\ p\frac{S_2}{N_1}c_{21} & p\frac{S_2}{N_2}c_{22} - \gamma \end{pmatrix} \tag{13}$$

which we can decompose as $J = T + \Sigma$ as follows

$$\begin{pmatrix} p\frac{S_1}{N_1}c_{11} - \gamma & p\frac{S_1}{N_2}c_{12} \\ p\frac{S_2}{N_1}c_{21} & p\frac{S_2}{N_2}c_{22} - \gamma \end{pmatrix} = \begin{pmatrix} -\gamma & 0 \\ 0 & -\gamma \end{pmatrix} + \begin{pmatrix} p\frac{S_1}{N_1}c_{11} & p\frac{S_1}{N_2}c_{12} \\ p\frac{S_2}{N_1}c_{21} & p\frac{S_2}{N_2}c_{22} \end{pmatrix}\ . \tag{14}$$

Notice that the left matrix in the decomposition is, indeed, a description of transitions: the flow of infected people into recovery. And, the right matrix in the decomposition is, indeed, a description of transmission: the creation of new infections through the contacts between S_1, S_2, I_1, and I_2.

Given this decomposition, the NGM is then given by $G = -T\Sigma^{-1}$. We thus get an NGM of

$$G = \begin{pmatrix} \frac{S_1}{N_1}\frac{pc_{11}}{\gamma} & \frac{S_1}{N_2}\frac{pc_{12}}{\gamma} \\ \frac{S_2}{N_1}\frac{pc_{21}}{\gamma} & \frac{S_2}{N_2}\frac{pc_{22}}{\gamma} \end{pmatrix} \tag{15}$$

And, because our goal is to understand R_0, we'll evaluate G at the disease-free equilibrium, where the population is (as assumed by R_0 logic) entirely susceptible. This means that $S_1 = N_1$ and $S_2 = N_2$. Thus,

$$G_0 = \begin{pmatrix} \frac{N_1}{N_1}\frac{pc_{11}}{\gamma} & \frac{N_1}{N_2}\frac{pc_{12}}{\gamma} \\ \frac{N_2}{N_1}\frac{pc_{21}}{\gamma} & \frac{N_2}{N_2}\frac{pc_{22}}{\gamma} \end{pmatrix}\ . \tag{16}$$

How can we interpret this equation? Think back to our initial interpretation of R_0: if we started with $I(0) = 1$ infection, we'd have $I(1) = R_0$ in the next generation, $I(2) = R_0^2$ in the generation after that, and then $I(3) = R_0^3$, and so on, provided that the population remains, to a good approximation, entirely susceptible.[5] In other words, early on in the epidemic, the infected subsystem grows with each generation like $I(\tau) \sim R_0^\tau$, where τ is the generation number.

[5] Not coincidentally, this is the same regime where the linearized dynamics are a good approximation of the true dynamics.

Here, instead, we have G_0, which tells us something similar: it is a matrix whose entries tell us how the elements of our infected subsystem grow from one generation of infections to the next. In other words, if $\mathbf{I}(0)$ represents some initial infections (e.g. $\mathbf{I}(0) = [1, 1]$), then $\mathbf{I}(1) = G_0\,\mathbf{I}(0)$, $\mathbf{I}(2) = G_0^2\,\mathbf{I}(0)$, and in general,

$$\mathbf{I}(\tau) = G_0^\tau\,\mathbf{I}(0)$$

In other words, in every generation, the vector of infection counts $\mathbf{I}$ will grow or shrink depending on what happens when it is multiplied by G_0, and mathematically, this is given by the **largest eigenvalue** of that matrix. How do we know? See the previous section, and compare the equation above to Equation (6)!

As a result, R_0 is given by the largest eigenvalue of the NGM evaluated at the disease free equilibrium, G_0. In the case of two populations of size $N_1 = N_2 = 1000$, with $p = \gamma = 1$, we get[6]

$$G_0 = \begin{pmatrix} c_{11} & c_{12} \\ c_{21} & c_{22} \end{pmatrix} \qquad \Longrightarrow \qquad R_0 = \frac{(c_{11} + c_{22}) + \sqrt{(c_{11} + c_{22})^2 - 4\,(c_{11}c_{22} - c_{12}c_{21})}}{2}$$

In fact, by calculating R_0 for our three examples in Figure 1, we get $R_0 = 1.5$, $R_0 = 1.52$, and $R_0 = 1.51$. In other words, in ever scenario, $R_0 > 1$! This explains why, even when we markedly decreased the within-group contact rate for group two, the overall system nevertheless exhibited an epidemic.

6.4 Summary and Generalization

We now generalize what we have learned above. First, let's imagine that instead of everyone being equally susceptible to infection (equal p), each group instead has its own p_m, with $\mathbf{p}$ representing the susceptibilities in vector form. Similarly, let's imagine that instead of everyone recovering at an equal rate, each group instead has its own γ_m, written as γ in vector form. Our equations are then

$$\begin{aligned} \dot{\mathbf{s}} &= -\left(D_{\mathbf{s}} D_{\mathbf{p}} C D_\omega^{-1}\right) \mathbf{i} \\ \dot{\mathbf{i}} &= \left(D_{\mathbf{s}} D_{\mathbf{p}} C D_\omega^{-1}\right) \mathbf{i} - D_\gamma \mathbf{i} \\ \dot{\mathbf{r}} &= D_\gamma \mathbf{i} \end{aligned} \tag{17}$$

and our NGM is

$$M = \left(D_{\mathbf{p}} D_s C D_\omega^{-1}\right) D_\gamma^{-1} \tag{18}$$

with

$$R_0 = \text{largest eigenvalue of } M \text{ setting } s = . \tag{19}$$

We can also draw a flow diagram for this system. Here, it helps to condense multiple terms into a single force of infection λ_m for each group m, as shown in Fig. 2.

[6] How? Recall that for a 2 × 2 matrix, the eigenvalues are $\lambda_{1,2} = \frac{tr \pm \sqrt{tr^2 - 4det}}{2}$, where tr is the trace and det is the determinant. Here, we want the larger of the two eigenvalues and thus take the $+$ from the $\pm$.

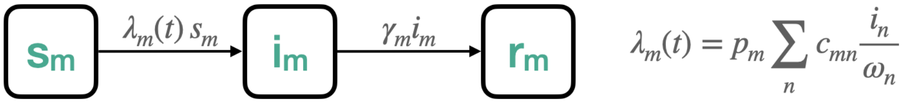

Figure 2: A flow diagram for the SIR model written in terms of subpopulation m, with susceptibility p_m, recover rate γ_m, contact matrix C, and population proportion $\omega_m = N_m/N$ (where $N = \sum_m N_m$). Notice that this model is structurally identical to the SIR model, but now simply includes variability in how each group experiences exposure and infection, and include an arbitrary contact matrix.

The NGM approach is powerful, and may be more complicated when the structure of the system is more complicated than a simple SIR framework. Still, this approach works generically![7]

6.5 Contact structures in real populations

In 2008, Mossong et al. published a paper in which 7,290 participants recorded 97,904 physical contacts over the course of one day. Those participants recorded the age, sex, location, duration, frequency, and occurrence of contacts. The researchers then assembled this information into population mixing matrices, i.e. empirical estimates of what the C matrix might be, based on various characteristics. Empirical estimates were updated in the COVID-19 era by Prem et al. and were also smoothed using a model.[8] Let's look at some real contact matrices (Figure 3).

There are a few key things to observe in real household structure contact matrices when we sort by age. First, we can see a dominant diagonal structure. This tells us that contact patterns are highly **assortative** by age, meaning that people interact with others whose age is similar to their own. Second, we can see two clear off-diagonal bands, approximately 30 years offset. These are patterns of parents and children.[9] Third, notice that when comparing the matrices in Fig. 3 we can see that the overall levels of contact are not identical, and that, for instance, the BE matrix shows more assortative contact structure among the elderly when compared with the PL matrix. This hints to us that BE may be demographically older.

Finally, we can see that these contact matrices are *non-symmetric* matrices, meaning that $c_{mn} \neq c_{nm}$, i.e., $C \neq C^T$. The reason for this is that c_{mn} reports from the perspective of someone in group m, and tells us the number of n-aged people they come into contact with, per day. In contrast, c_{nm} reports on the same contacts, but from the perspective of someone in group n. These two numbers need not be the same, which we can illustrate with a simple example: imagine a room with 10 children and 2 adults in it, and everyone contacts everyone, meaning that there were 20 total contacts. The 10 children have contacts 2 adults each, but the 2 adults have contacted 10 children each. In principle, if we assume that any physical contact

[7] See Diekmann, Heesterbeek, and Roberts, 2009 for a more in-depth explanation with examples.

[8] This means that the smoothing wasn't just local averaging. In this case the authors used prior knowledge about contact matrices in a Bayesian approach to smoothing.

[9] You may also notice a second band offset by arounn 60 years, representing grandparents and grandchildren!

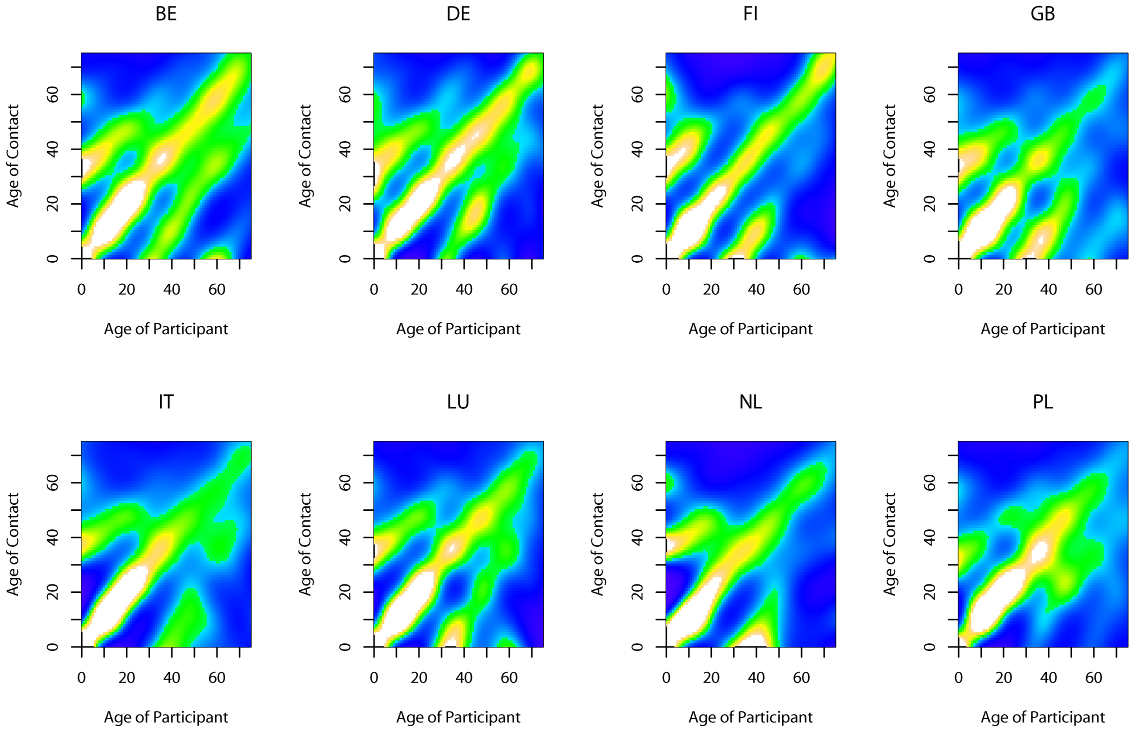

Figure 3: Contact matrices from Mossong et al (Fig 3B). Each panel corresponds to a European country: BE, Belgium; DE, Germany; FI, Finland; GB, Great Britain; IT, Italy; LU, Luxembourg; NL, The Netherlands; PL, Poland. What do you notice is similar or different between these? How might we consider targeted interventions differently in these two populations? What additional information might you need to know to target interventions well?

between two people counts as a contact for each of them, then the total contacts must balance, meaning that $\omega_m c_{mn} = \omega_n c_{nm}$. However, this need not be the case for empirically derived contact patterns that reflect contact diaries from people's experiences.

6.6 Outlook

Models with contact structure imagine that, when we put people into compartments, those compartments form a sort of grid: on one axis we have the labels of our model's structure, such as S, I, and R, and on the other axis we have the labels of our population's groups, such as age. This means that we can use the same model structure but over different population groups, or we can have a different model structure (perhaps inclusive of vaccination), but over the same population groups. In general, we are free to combine model structures and population groups, depending on the task at hand.

One thing you may have noticed is that the only interaction between groups in our examples here was that people of different ages *come into contact*, but we did not allow people to move between ages. These types of models exist too, in which individuals are born into the youngest age group, and over time age into older compartments. Such models add some complexity, but are a good choice when the dynamics of transmission and immunity are on a similar timescale as the dynamics of aging.