From the book *Networks, Crowds, and Markets: Reasoning about a Highly Connected World.*
By David Easley and Jon Kleinberg. Cambridge University Press, 2010.
Complete preprint on-line at http://www.cs.cornell.edu/home/kleinber/networks-book/

Chapter 21

Epidemics

The study of epidemic disease has always been a topic where biological issues mix with social ones. When we talk about epidemic disease, we will be thinking of contagious diseases caused by biological pathogens — things like influenza, measles, and sexually transmitted diseases, which spread from person to person. Epidemics can pass explosively through a population, or they can persist over long time periods at low levels; they can experience sudden flare-ups or even wave-like cyclic patterns of increasing and decreasing prevalence. In extreme cases, a single disease outbreak can have a significant effect on a whole civilization, as with the epidemics started by the arrival of Europeans in the Americas [130], or the outbreak of bubonic plague that killed 20% of the population of Europe over a seven-year period in the 1300s [293].

21.1 Diseases and the Networks that Transmit Them

The patterns by which epidemics spread through groups of people is determined not just by the properties of the pathogen carrying it — including its contagiousness, the length of its infectious period, and its severity — but also by network structures within the population it is affecting. The social network within a population — recording who knows whom — determines a lot about how the disease is likely to spread from one person to another. But more generally, the opportunities for a disease to spread are given by a *contact network*: there is a node for each person, and an edge if two people come into contact with each other in a way that makes it possible for the disease to spread from one to the other.

This suggests that accurately modeling the underlying network is crucial to understanding the spread of an epidemic. This has led to research studying how travel patterns within a city [149, 295] or via the worldwide airline network [119] could affect the spread of a fast-moving disease. Contact networks are also important in understanding how diseases spread through animal populations — with researchers tracing out the interactions within livestock populations during epidemics such as the 2001 foot-and-mouth outbreak in the United Kingdom [211] — as well as plant populations, where the affected individuals occupy fixed locations and diseases tend to have a much clearer spatial footprint [139]. And similar models have been employed for studying the spread of computer viruses, with malicious software spreading between computers across an underlying communication network [241].

The pathogen and the network are closely intertwined: even within the same population, the contact networks for two different diseases can have very different structures, depending on the diseases' respective modes of transmission. For a highly contagious disease, involving airborne transmission based on coughs and sneezes, the contact network will include a huge number of links, including any pair of people who sat together on a bus or an airplane. For a disease requiring close contact, or a sexually transmitted disease, the contact network will be much sparser, with many fewer pairs of people connected by links. Similar distinctions arise in studying computer viruses, where a piece of software infecting computers across the Internet will have a much broader contact network than one that spreads by short-range wireless communication between nearby mobile devices [251].

Connections to the Diffusion of Ideas and Behaviors. There are clear connections between epidemic disease and the diffusion of ideas through social networks. Both diseases and ideas can spread from person to person, across similar kinds of networks that connect people, and in this respect, they exhibit very similar structural mechanisms — to the extent that the spread of ideas is often referred to as "social contagion" [85]. Having considered the diffusion of ideas, innovations, and new behaviors in Chapter 19, why then are we revisiting this topic afresh in the context of diseases?

In the context of our discussions here about networks, the biggest difference between biological and social contagion lies in the process by which one person "infects" another. With social contagion, people are making decisions to adopt a new idea or innovation, and our models in Chapter 19 were focused on relating the underlying decision-making processes to the larger effects at the network level. With diseases, on the other hand, not only is there a lack of decision-making in the transmission of the disease from one person to another, but the process is sufficiently complex and unobservable at the person-to-person level that it is most useful to model it as *random*. That is, we will generally assume that when two people are directly linked in the contact network, and one of them has the disease, there is a given probability that he or she will pass it to the other. This use of randomness allows us to abstract away questions about the mechanics of how one person catches a disease from another for which we have no useful simple models.

This, then, will be the concrete difference in our discussion of biological as opposed to social contagion — not so much the new context as the new classes of models, based on random processes in networks, that will be employed. In the next three sections, we discuss some of the most basic probabilistic models for epidemics in networks; we then consider how these models provide insight into some basic qualitative issues in the spread of disease, including synchronization, timing, and concurrency in transmission. Finally, we discuss how some of the models developed here are related to similar issues in genetic inheritance, where a kind of randomized propagation takes place through genealogical networks.

Before moving on to this, it is worth noting that randomized models can also sometimes be useful in studying social contagion, particularly in cases where the underlying decision processes of the individuals are hard to model and hence more usefully abstracted as random. Often the two approaches — decision-based and probabilistic — produce related results, and they can sometimes be used in conjunction [62, 408]. Understanding the relationship between these methodologies at a deeper level is an interesting direction for further research.

21.2 Branching Processes

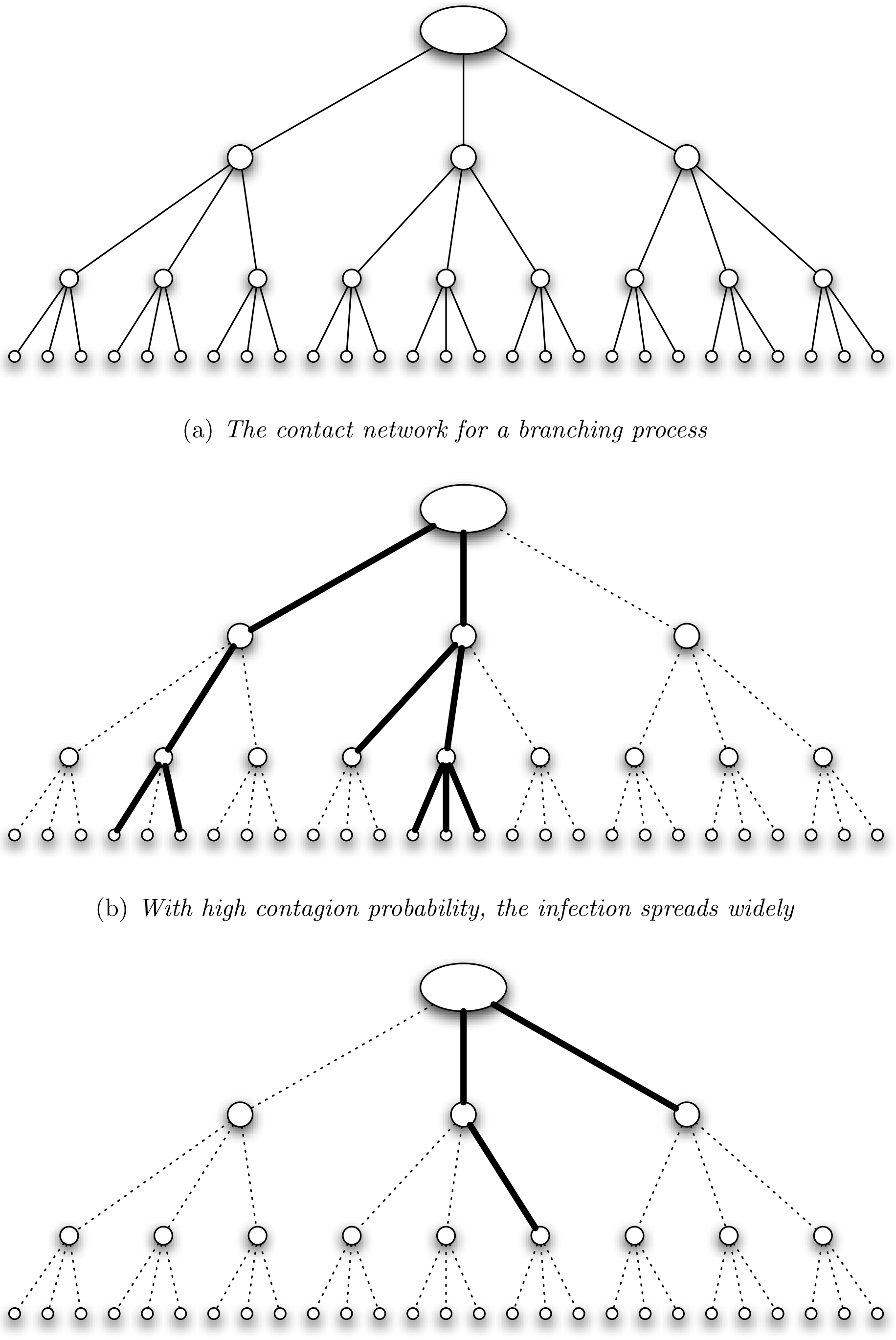
We begin with perhaps the simplest model of contagion, which we refer to as a *branching process*. It works as follows.

- *(First wave.)* Suppose that a person carrying a new disease enters a population, and transmits it to each person he meets independently with a probability of p. Further, suppose that he meets k people while he is contagious; let's call these k people the *first wave* of the epidemic. Based on the random transmission of the disease from the initial person, some of the people in the first wave may get infected with the disease, while others may not.

- *(Second wave.)* Now, each person in the first wave goes out into the population and meets k different people, resulting in a *second wave* of $k \cdot k = k^2$ people. Each infected person in the first wave passes the disease independently to each of the k second-wave people they meet, again independently with probability p.

- *(Subsequent waves.)* Further waves are formed in the same way, by having each person in the current wave meet k new people, passing the disease to each independently with probability p.

Thus the contact network for this epidemic can be drawn as in Figure 21.1(a) (with $k = 3$ and only the first three waves shown). We refer to such a network as a *tree*: it has a single node at the top called the *root*; every node is connected to a set of nodes in the level below it; and every node but the root is also connected to a *single* node in the level above it. The tree that forms the contact network for the branching process is in fact infinite, since we continue defining waves indefinitely.



(a) *The contact network for a branching process*

(b) *With high contagion probability, the infection spreads widely*

(c) *With low contagion probability, the infection is likely to die out quickly*

Figure 21.1: The branching process model is a simple framework for reasoning about the spread of an epidemic as one varies both the amount of contact among individuals and the level of contagion.

Now, what is the behavior of an epidemic in this model? We can picture the spread of the epidemic by highlighting the edges of the contact network on which the disease passes successfully from one person to another — recall that each of these infections happens independently with probability p. Thus, Figure 21.1(b) shows an aggressive epidemic that infects two people in the first wave, three in the second wave, five in the third wave, and presumably more in future waves (not shown in the picture). Figure 21.1(c), on the other hand, shows a much milder epidemic (for a less contagious disease, with a smaller value of p): of the two people infected in the first wave, one doesn't infect anyone else, and the other infects only one further person who in turn doesn't pass it on. This disease has completely vanished from the population after the second wave, having infected only four people in total.

The Basic Reproductive Number and a Dichotomy for Branching Processes. Our last observation about Figure 21.1(c) reflects a fundamental property of branching processes: if the disease in a branching process ever reaches a wave where it fails to infect anyone, then it has died out: since people in future waves can only catch the disease from others higher up in the tree, no one in any future wave will be infected either.

So there are really only two possibilities for a disease in the branching process model: it reaches a wave where it infects no one, thus dying out after a finite number of steps; or it continues to infect people in *every* wave, proceeding infinitely through the contact network. And it turns out that there is a simple condition to tell these two possibilities apart, based on a quantity called the *basic reproductive number* of the disease.

The basic reproductive number, denoted R_0, is the expected number of new cases of the disease caused by a single individual. Since in our model everyone meets k new people and infects each with probability p, the basic reproductive number here is given by $R_0 = pk$. The outcome of the disease in a branching process model is determined by whether the basic reproductive number is smaller or larger than 1.

> *Claim: If $R_0 < 1$, then with probability 1, the disease dies out after a finite number of waves. If $R_0 > 1$, then with probability greater than 0 the disease persists by infecting at least one person in each wave.*

We give a proof of this claim in Section 21.8. Even without the details of the proof, however, we can see that the basic condition expressed in the claim — comparing R_0 to 1 — has a natural intuitive basis. When $R_0 < 1$, the disease isn't able to replenish itself: each infected person produces less than one new case in expectation, and so — even if it grows briefly due to the outcome of random fluctuations — the size of the outbreak is constantly trending downward. When $R_0 > 1$, on the other hand, the size of outbreak is constantly trending upward. Notice, however, that even when $R_0 > 1$, the conclusion is simply that the disease persists with positive probability, not with absolute certainty: whenever $p < 1$, then there is always some chance that none of the first few infected people will succeed in infecting anyone else, causing the disease to die out. In other words, even an ultra-contagious disease can simply get "unlucky" and vanish from the population before it has a chance to really get going.

The dichotomy expressed by this condition has an interesting "knife-edge" quality to it when R_0 is close to 1. In particular, suppose we have a branching process where R_0 is very slightly below 1, and we increase the contagion probability p by a little bit; the result could push R_0 above 1, suddenly resulting in a positive probability of an enormous outbreak. The same effect can happen in the reverse direction as well, where slightly reducing the contagiousness of a disease to push R_0 below 1 can eliminate the risk of a large epidemic. And since R_0 is the product of p and k, small changes in the number of people k that each person comes into contact with can also have a large effect when R_0 is near 1.

All this suggests that around the critical value $R_0 = 1$, it can be worth investing large amounts of effort even to produce small shifts in the basic reproductive number. Since R_0 is the product of the two terms p and k, it is in fact easy to interpret two basic kinds of public-health measures in terms of reductions to R_0: quarantining people, which reduces the quantity k, and encouraging behavioral measures such as better sanitary practices to reduce the spread of germs, which reduces the quantity p.

The branching process model is clearly a very simplified model of disease-spreading; the structure of the contact network, with no triangles at all, is reminiscent of our first pass at a model for the small-world phenomenon in Chapter 20. Thus, in the next few sections, we will look at models that can handle more complex contact networks. For these models, a dichotomy as simple as the one in the Claim above does not hold. However, the notion of the basic reproductive number is still a useful heuristic guide to the behavior of more complex models; even when epidemiological modelers do not have a precise condition governing when an epidemic will persist and when it will die out, they find the reproductive number R_0 to be a useful approximate indication of the spreading power of the disease.

21.3 The SIR Epidemic Model

We now develop an epidemic model that can be applied to any network structure. To do this, we preserve the basic ingredients of the branching process model at the level of individual nodes, but make the contact structure much more general. An individual node in the branching process model goes through three potential stages during the course of the epidemic:

- *Susceptible:* Before the node has caught the disease, it is susceptible to infection from its neighbors.

- *Infectious:* Once the node has caught the disease, it is infectious and has some probability of infecting each of its susceptible neighbors.

- *Removed:* After a particular node has experienced the full infectious period, this node is removed from consideration, since it no longer poses a threat of future infection.

Using this three-stage "life cycle" for the disease at each node, we now define a model for epidemics on networks. We are given a directed graph representing the contact network; so an edge pointing from v to w in the graph means that if v becomes infected at some point, the disease has the potential to spread directly to w. To represent a symmetric contact between people, where either has the potential to directly infect the other, we can put in directed edges pointing each way: both from v to w and also from w to v. Since contacts between people are often symmetric, it is fine to use networks where most edges appear in each direction, but it is sometimes convenient to be able to express asymmetric contacts as well.

Now, each node has the potential to go through the Susceptible-Infectious-Removed cycle, where we abbreviate these three states as S, I, and R. The progress of the epidemic is controlled by the contact network structure and by two additional quantities: p (the probability of contagion) and t_I (the length of the infection).

- Initially, some nodes are in the I state and all others are in the S state.

- Each node v that enters the I state remains infectious for a fixed number of steps t_I.

- During each of these t_I steps, v has a probability p of passing the disease to each of its susceptible neighbors.

- After t_I steps, node v is no longer infectious or susceptible to further bouts of the disease; we describe it as *removed* (R), since it is now an inert node in the contact network that can no longer either catch or transmit the disease.

This describes the full model; we refer to it as the *SIR model*, after the three disease states that nodes experience. Figure 21.2 shows an example of the SIR model unfolding on a particular contact network through successive steps; in each step, shaded nodes with dark borders are in the I state and shaded nodes with thin borders are in the R state.

The SIR model is clearly most appropriate for a disease that each individual only catches once in their lifetime; after being infected, a node is removed either because it has acquired lifetime immunity or because the disease has killed it. In the next section, we will consider a related model for diseases that can be caught multiple times by the same person. Notice also that the branching process model from Section 21.2 is a special case of the SIR model: it simply corresponds to the SIR model where $t_I = 1$ and the contact network is an infinite tree, with each node connected to a fixed number of neighbors in the level below.

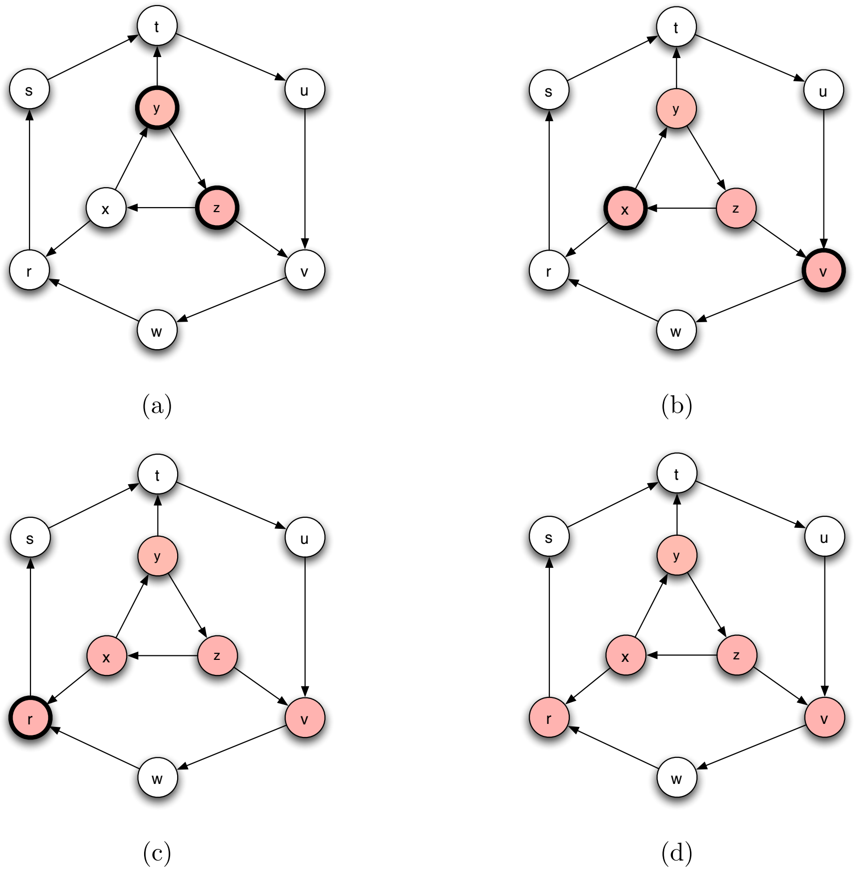


Figure 21.2: The course of an SIR epidemic in which each node remains infectious for a number of steps equal to $t_I = 1$. Starting with nodes y and z initially infected, the epidemic spreads to some but not all of the remaining nodes. In each step, shaded nodes with dark borders are in the Infectious (I) state and shaded nodes with thin borders are in the Removed (R) state.

Extensions to the SIR model. Although the contact network in the general SIR model can be arbitrarily complex, the disease dynamics are still being modeled in a simple way. Contagion probabilities are set to a uniform value p, and contagiousness has a kind of "on-off" property: a node is equally contagious for each of the t_I steps while it has the disease.

However, it is not difficult to extend the model to handle more complex assumptions. First, we can easily capture the idea that contagion is more likely between certain pairs of nodes by assigning a separate probability $p_{v,w}$ to each pair of nodes v and w for which v links to w in the directed contact network. Here, higher values of $p_{v,w}$ correspond to closer contact and more likely contagion, while lower values indicate less intensive contact. We can also choose to model the infectious period as random in length, by assuming that an infected node has a probability q of recovering in each step while it is infected, while leaving

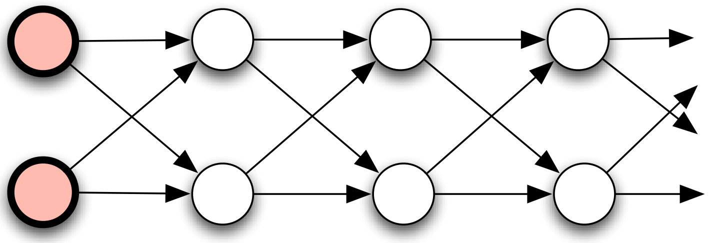


Figure 21.3: In this network, the epidemic is forced to pass through a narrow "channel" of nodes. In such a structure, even a highly contagious disease will tend to die out relatively quickly.

the other details of the model as they are.

More elaborate extensions to the model involve separating the I state into a sequence of several states (e.g. early, middle, and late periods of the infection), and allowing the contagion probabilities to vary across these states [238]. This could be used, for example, to model a disease with a highly contagious incubation period, followed by a less contagious period while symptoms are being expressed. Researchers have also considered variations on the SIR model in which the disease-causing pathogen is mutating (and thus changing its disease characteristics) over the course of the outbreak [183].

The Role of the Basic Reproductive Number. We now discuss some observations about the SIR model, focusing on the most basic version of the model in an arbitrary network. First, let's recall the claim made at the end of Section 21.2, that in networks that do not have a tree structure, the simple dichotomy in epidemic behavior determined by the basic reproductive number R_0 does not necessarily hold. In fact, it is not hard to construct an example showing how this dichotomy breaks down. To do this, let's start with the network depicted in Figure 21.3, and suppose that these layers of two nodes at a time continue indefinitely to the right. Let's consider an SIR epidemic in which $t_I = 1$, the infection probability p is $2/3$, and the two nodes at the far left are the ones that are initially infected.

When we don't have a tree network, we need to decide how to define an analogue of the basic reproductive number. In a network as highly structure as the one in Figure 21.3, we can work directly from the definition of R_0 as the expected number of new cases of the disease caused by a single individual. (For less structured networks, one can consider R_0 to be the expected number of new cases caused by a randomly chosen individual from the population.) In Figure 21.3, each infected node has edges to two nodes in the next layer; since it infects each with probability $2/3$, the expected number of new cases caused by this

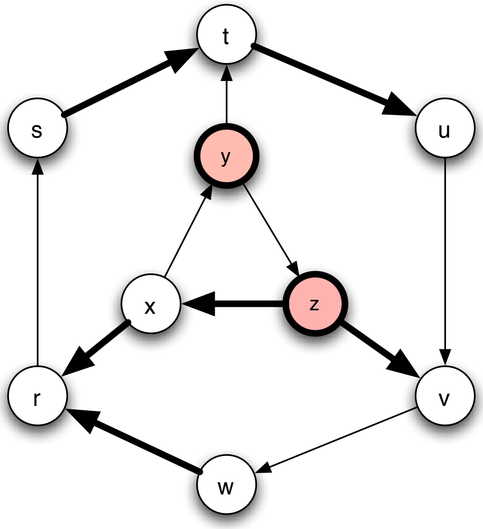


Figure 21.4: An equivalent way to view an SIR epidemic is in terms of *percolation*, where we decide in advance which edges will transmit infection (should the opportunity arise) and which will not.

node is 4/3.

So in our example, $R_0 > 1$. Despite this, however, it is easy to see that the disease will die out almost surely after reaching only a finite number of steps. In each layer, there are four edges leading to the next layer, and each will independently fail to transmit the disease with probability 1/3. Therefore, with probability $(1/3)^4 = 1/81$, all four edges will fail to transmit the disease — and at this point, these four edges become a “roadblock” guaranteeing the disease can never reach the portion of the network beyond them. Thus, as the disease moves along layer-by-layer, there is a probability of at least 1/81 that each layer will be its last. Therefore, with probability 1, it must come to an end after a finite number of layers.

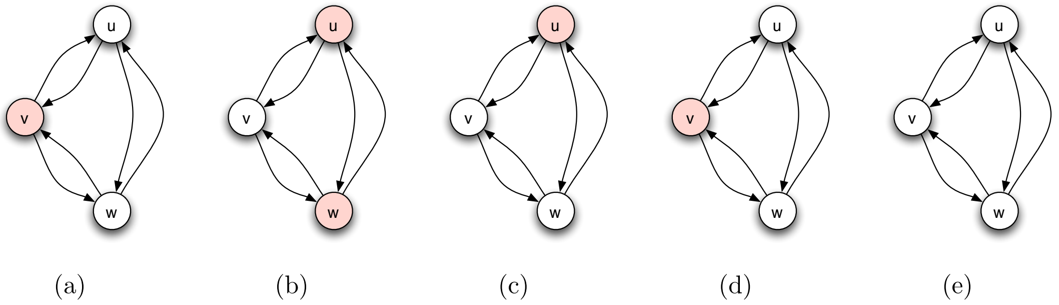
This is a very simple example, but it already indicates how different network structures can be more or less conducive to the spread of a disease — even taking contagiousness and other disease properties as given. Whereas the contact network of the simple branching process from Section 21.2 was a tree that expanded rapidly in all directions, the network in Figure 21.3 forces the disease to pass through a narrow “channel” in which a small breakdown in contagion can wipe it out. Understanding how specific types of network structure interact with disease dynamics remains a challenging research question, and one that affects predictions about the course of real epidemics.

SIR Epidemics and Percolation. Thus far we have been thinking about SIR epidemics as dynamic processes, in which the state of the network evolves step-by-step over time. This captures the temporal dynamics of the disease itself as it spreads through a population. Interestingly, however, there is an equivalent and completely static view of these epidemics that is often very useful from a modeling point of view [44, 173].

We now describe how to arrive at this static view of the process, focusing on the basic SIR model in which $t_I = 1$. Consider a point in an SIR epidemic when a node v has just become infectious, and it has a susceptible neighbor w. Node v has one chance to infect w (since $t_I = 1$), and it succeeds with probability p. We can view the outcome of this random event as being determined by flipping a coin that has a probability p of coming up *"heads"*, and observing the outcome. From the point of view of the process, it clearly does not matter whether the coin was flipped at the moment that v first became infectious, or whether it was flipped at the very beginning of the whole process and is only being revealed now. Continuing this reasoning, we can in fact assume that for *each* edge in the contact network — from a node v to a node w — a coin with *"heads"* probability p is flipped at the very beginning of the process (independently of the coins for all other pairs of neighbors), and the result is stored so that it can be later checked *in the event* that v becomes infectious while w is susceptible.

With all the coins flipped in advance, the SIR process can be viewed as follows. The edges in the contact network for which the coin flip is successful are declared to be *open*; the remaining edges are declared to be *blocked*. The situation is now as pictured in Figure 21.4, which shows a sample result of coin flips consistent with the pattern of infections in the example from Figure 21.2. And we can now see how to make use of the open and blocked edges to represent the course of the epidemic: *A node v will become infected during the epidemic if and only if there is a path to v from one of the initially infected nodes that consists entirely of open edges.* Thus, while Figure 21.4 looks superficially different from the sequence of stages in Figure 21.2, it is in fact a beautifully compact way to summarize the course of the epidemic: the nodes that are eventually infected are precisely those that can be reached from the initially infected nodes along a sequence of open edges in the network.

This static view of the model is often referred to as *percolation*, due to the following physical analogy. If we think of the contact network as a system of pipes, and the pathogen as a fluid moving through these pipes, then the edges in the contact network on which contagion succeeds are the "open" pipes and the edges on which it fails are the "blocked" pipes. We now want to know which nodes the fluid will reach, given that it can only pass through open pipes. In fact, this is not simply an illustrative metaphor; percolation is a topic that has been extensively studied by physicists and mathematicians as a model for the flow of fluids through certain types of porous media [69, 173]. It is both an interesting topic in its own right, and useful for its role as an equivalent view of the progress of an epidemic.



(a) (b) (c) (d) (e)

Figure 21.5: In an SIS epidemic, nodes can be infected, recover, and then be infected again. In each step, the nodes in the Infectious state are shaded.

21.4 The SIS Epidemic Model

In the previous sections we have been considering models for epidemics in which each individual contracts the disease at most once. However, a simple variation on these models allows us to reason about epidemics where nodes can be reinfected multiple times.

To represent such epidemics, we have nodes that simply alternate between two possible states: *Susceptible (S)* and *Infectious (I)*. There is no *Removed* state here; rather, after a node is done with the Infectious state, it cycles back to the Susceptible state and is ready to catch the disease again. Because of this alternation between the S and I states, we refer to the model as the *SIS model*.

Aside from the lack of an R state, the mechanics of the model follow the SIR process very closely.

- Initially, some nodes are in the I state and all others are in the S state.

- Each node v that enters the I state remains infectious for a fixed number of steps t_I.

- During each of these t_I steps, v has a probability p of passing the disease to each of its susceptible neighbors.

- After t_I steps, node v is no longer infectious, and it returns to the S state.

Figure 21.5 shows an example of the SIS model unfolding on a three-node contact network with $t_I = 1$. Notice how node v starts out infected, recovers, and later becomes infected again — we can imagine this as the contact network within a three-person apartment, or a three-person family, where people pass a disease on to others they're living with, and then get it back from them later.

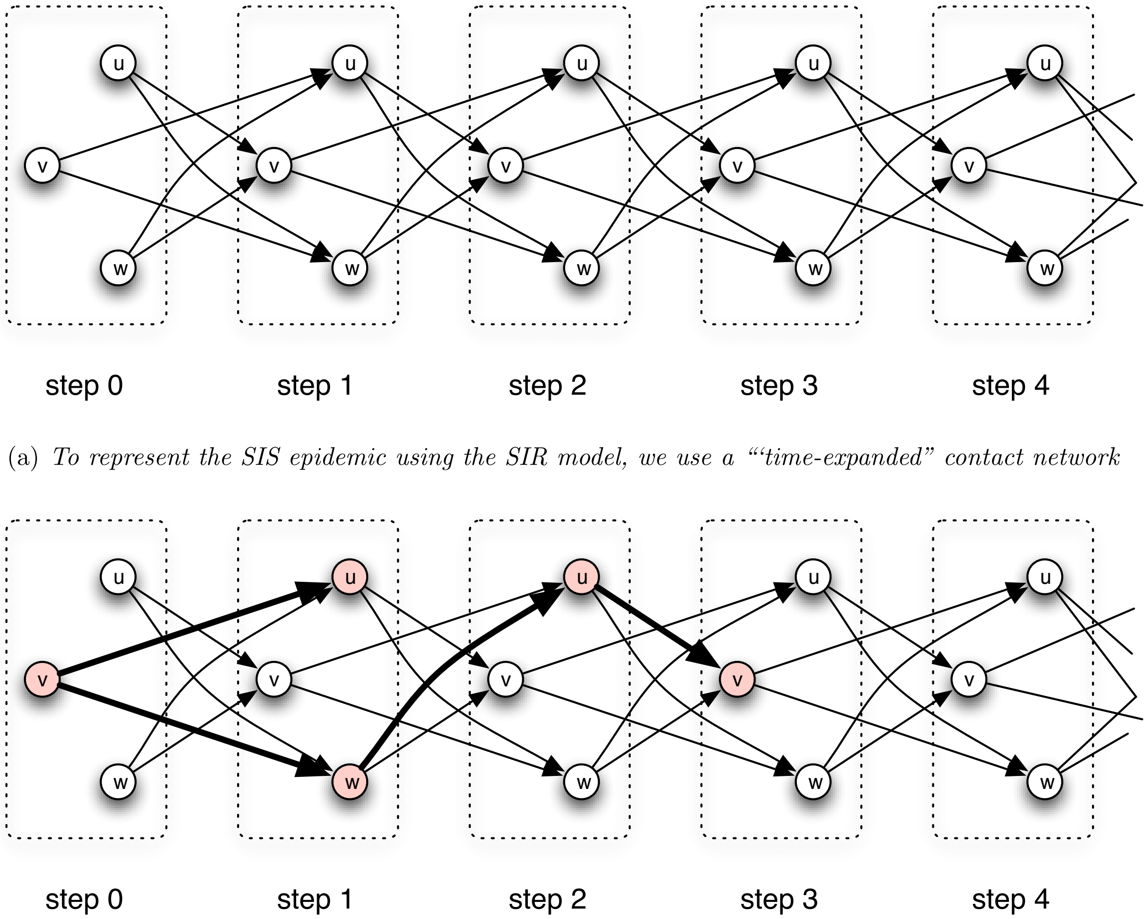
As with the SIR model, the SIS model can be extended to handle more general kinds of assumptions: different contagion probabilities between different pairs of people; probabilistic recovery from the disease, in which each infected node transitions back to the susceptible state with probability q each step; and multiple stages of infection, with varying disease properties across them.

Life Cycles of SIR and SIS Epidemics. The examples in this section and preceding one suggest that the overall "trajectories" of SIR and SIS epidemics on (finite-size) graphs are qualitatively quite different. An SIR epidemic on a finite graph is burning through a bounded supply of nodes — since nodes can never be reinfected — and therefore it must come to an end after a relatively small number of steps. An SIS epidemic, on the other hand, can run for an extremely long time as it cycles through the nodes potentially multiple times. But as Figure 21.5(e) illustrates, if there ever comes a point in an SIS epidemic when all nodes are simultaneously free of the disease, then the epidemic has died forever: there are no longer any infected individuals to pass the disease to others. And on a finite graph, there will eventually (with probability 1) come a point in time when all contagion attempts simultaneously fail for t_I steps in a row, and at this point it will be over. Thus a key question with an SIS epidemic on a given contact network is to understand how long the outbreak will last, and how many individuals will be affected at different points in time.

For contact networks where the structure is mathematically tractable, researchers have in fact proved "knife-edge" results for the SIS model similar to our dichotomy for branching processes. These results, on particular classes of contact networks, show that at a particular critical value of the contagion probability p, an SIS epidemic on the network will undergo a rapid shift from one that dies out quickly to one that persists for a very long time [52, 278]. This type of analysis tends to be mathematically quite complex, with this critical value of the contagion probability p depending in subtle ways on the structure of the network.

A Connection Between SIR and SIS Epidemics. Despite the differences between the SIR and SIS models, in fact it is possible to represent some of the basic variants of the SIS model as special cases of the SIR model. This surprising relationship is further evidence of the flexibility of the basic epidemic models, in which formalisms defined in different ways turn out to have very close connections to each other.

We describe the relationship for the SIS model with $t_I = 1$, when each node is infectious for a single step before recovering. The key insight is that if we think about a node v as in fact being a "different individual" at each time step, then we can represent things so that nodes are never reinfected. Specifically, given an instance of the SIS model with $t_I = 1$, we create a separate copy of each node for each time step $t = 0, 1, 2, 3$ and onward. We will call this the *time-expanded contact network*. Now, for each edge in the original contact network, linking a node v to a node w, we create edges in the time-expanded contact network from the copy of v at time t to the copy of w at time $t + 1$; this simply encodes the idea that w



(a) *To represent the SIS epidemic using the SIR model, we use a "'time-expanded" contact network*

(b) *The SIS epidemic can then be represented as an SIR epidemic on this time-expanded network.*

Figure 21.6: An SIS epidemic can be represented in the SIR model by creating a separate copy of the contact network for each time step: a node at time t can infect its contact neighbors at time $t + 1$.

can potentially catch the disease at time $t + 1$ if v is infected at time t. Figure 21.6(a) shows this construction applied to the contact network from Figure 21.5.

The point is that the same SIS disease dynamics that previously circulated around in the original contact network can now flow forward in time through the time-expanded contact network, with copies of nodes that are in the I state at time t producing new infections in copies of nodes at time $t + 1$. But on this time-expanded graph we have an SIR process, since any copy of a node can be treated as removed (R) once its one time step of infection is over; and with this view of the process, we have the same distribution of outcomes as the original SIS process. Figure 21.6(b) shows the course of the SIR epidemic that corresponds to the SIS epidemic in Figure 21.5.

21.5 Synchronization

The models we've developed give us a framework for thinking about various broader issues in the spread of disease. We already encountered one of these issues in the dichotomy for branching processes, which provided a formal basis for the sensitivity of outbreaks to small variations in contagiousness, and for the crucial role of the basic reproductive number. We now look at a related issue in the global dynamics of a disease — the tendency of epidemics for certain diseases to synchronize across a population, sometimes producing strong oscillations in the number of affected individuals over time. Such effects are well-known for diseases including measles [196, 213] and syphilis [195].

When looking at public-health data, it is natural to look at periodic oscillations in the number of cases of a disease and to try positing external causes for the effect. For example, cycles in the prevalence of syphilis across the U.S. over the past 50 years have traditionally been attributed to large-scale societal changes, including changes in sexual mores and other forces [195]. While such factors clearly play a role, recent research has shown that oscillations and synchronization over time can in fact result largely from the contagion dynamics of the disease itself, and that similar patterns can be created in direct simulations of the disease using the types of models we have been considering here [195, 267].

We now describe how such effects can be produced using simple epidemic models. The crucial ingredients appear to be a combination of temporary immunity and long-range links in the contact network. Roughly, long-range links produce coordination in the timing of flare-ups across dispersed parts of the network; when these subside, the temporary immunity produces a network-wide deficit in the number and connectivity of susceptible individuals, yielding a large "trough" in the size of the outbreak that directly follows the "peak" from the earlier flare-ups. We now describe how to make this intuitive picture concrete using simple models.

The SIRS Epidemic Model. The first step in producing a model with oscillations is to allow the disease to confer temporary but not permanent immunity on infected individuals — a feature of many real diseases. To do this, we combine elements of the SIR and SIS models in a simple way, so that after an infected node recovers, it passes briefly through the R state on its way back to the S state. We call the resulting model the *SIRS model* [267], since nodes pass through the sequence S-I-R-S as the epidemic proceeds. In detail, the model works as follows.

- Initially, some nodes are in the I state and all others are in the S state.
- Each node v that enters the I state remains infectious for a fixed number of steps t_I.
- During each of these t_I steps, v has a probability p of passing the disease to each of its susceptible neighbors.

- *(The new feature of the model.)* After t_I steps, node v is no longer infectious. It then enters the R state for a fixed number of steps t_R. During this time, it cannot be infected with the disease, nor does it transmit the disease to other nodes. After t_R steps in the R state, node v returns to the S state.

For an SIRS epidemic, the course of the disease through a population is clearly affected not just by the quantities p and t_I, but also by the length t_R of the temporary immunity that is conferred.

Small-World Contact Networks. Temporary immunity can produce oscillations in very localized parts of the network, with patches of immunity following large numbers of infections in a concentrated area. But for this to produce large fluctuations that can be seen at the level of the full network, the flare-ups of the disease have to be coordinated so that they happen at roughly the same time in many different places. A natural mechanism to produce this kind of coordination is to have a network that is rich in long-range connections, linking otherwise far-apart sections of the network.

This kind of structure is familiar from our discussion of small-world properties in Chapter 20. There, we considered network models where many of the links were "local" and clustered — connecting nodes with very similar social and geographic characteristics, according to the principle of homophily — while some were long-range links, corresponding to weak ties that link very different parts of the network. In Chapter 20 we focused on the effect this kind of structure has on the distances between nodes. But there is a closely related consequence: long-range links make it possible for things that happen in one part of the network to quickly affect what is happening elsewhere.

Watts and Strogatz observed the relevance of small-world properties to synchronization in their original paper on the topic [411], and Kuperman and Abramson showed how it could naturally lead to synchronization and oscillation in epidemics [267]. For their analysis they constructed random networks with small-world properties, in a manner very similar to the grid-plus-random-edges construction discussed in Chapter 20; instead of the version from that chapter, they more closely followed the original construction of Watts and Strogatz, in which a ring network is rewired to produce random shortcuts [411]. Specifically, they started with a graph in which the nodes are arranged in a ring, and each node is connected to its neighbors for some number of steps in each direction. These are all homophilous links, in that they connect nodes that are very close together on the ring. Then, independently with some probability c, they turned each edge into a weak tie by rewiring one end of it to a node chosen uniformly at random. Thus, the probability c controls the fraction of links in the network that serve as long-range weak ties.

When the SIRS model is run on this kind of network, one finds very different behavior depending on the value of c, as indicated in Figure 21.7. When c is very small, disease

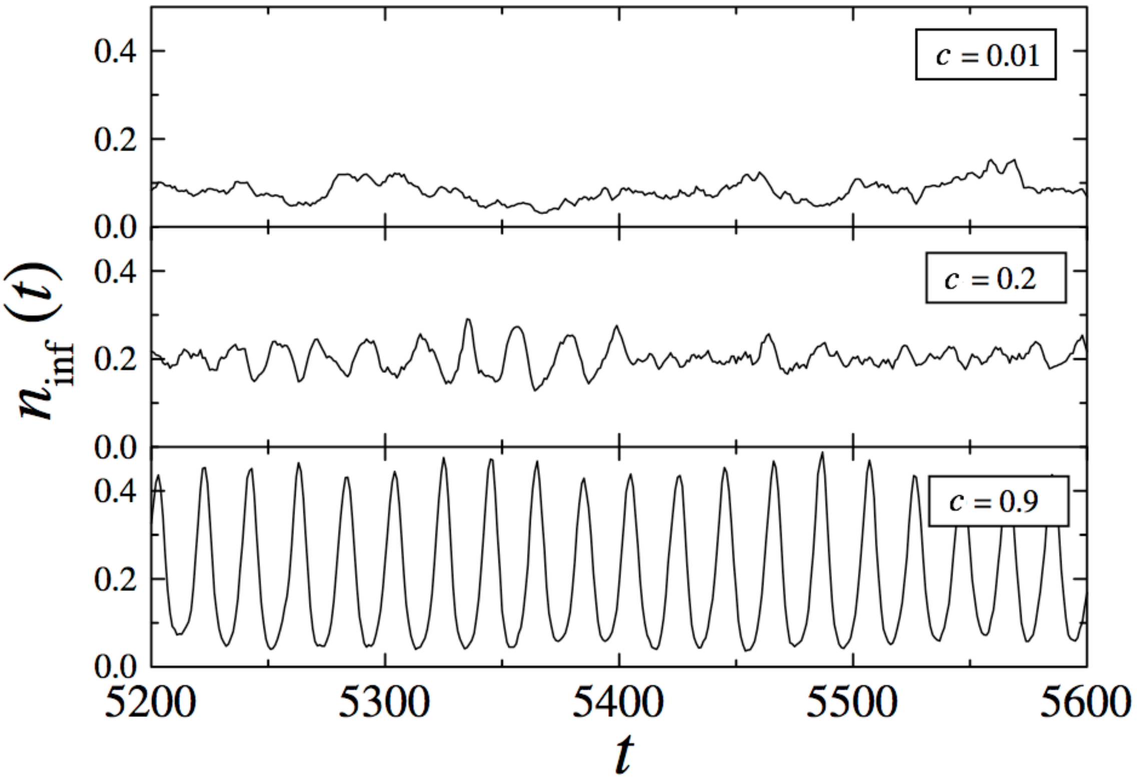


Figure 21.7: These plots depict the number of infected people over time (the quantity $n_{\text{inf}}(t)$ on the y-axis) by SIRS epidemics in networks with different proportions of long-range links. With c representing the fraction of long-range links, we see an abscence of oscillations for small c ($c = 0.01$), wide oscillations for large c ($c = 0.9$), and a transitional region ($c = 0.2$) where oscillations intermittently appear and then disappear. (Results and image from [267].)

transmission through the network occurs mainly via the short-range local edges, and so flare-ups of the disease in one part of the network never become coordinated with flare-ups in other parts. As c increases, these flare-ups start to synchronize, and since each burst produces a large number of nodes with temporary immunity, there is a subsequent trough as the disease has difficulty making its way through the sparser set of available targets. For very large values of c (such as $c = 0.9$ in Figure 21.7), there are clear waves in the number of affected individuals; for intermediate values of c (such as $c = 0.2$) one observes interesting effects in which the system achieves network-wide synchronization for a period, and then seems to fall back "out of sync" for reasons that are hard to quantify.

These results show how fairly complex epidemic dynamics can arise from simple models of contagion and contact structure. There are, however, a number of interesting open questions;

the results discussed here have been primarily found through simulation, and analyzing the onset of synchronization mathematically in this model remains largely unexplored.

Sychronization in Epidemic Data. It is possible to study these effects empirically — and evaluate proposed models — using extensive records of disease prevalence that reach back many years. Grassly, Fraser, and Garnett [195] performed an instructive comparison of syphilis and gonorrhea that illustrates a number of synchronization principles. The prevalence of syphilis exhibits prominent oscillations on an 8-11-year cycle, while gonorrhea exhibits very little in the way of periodic behavior. Yet the two diseases affect similar populations, and are presumably subject to very similar societal forces.

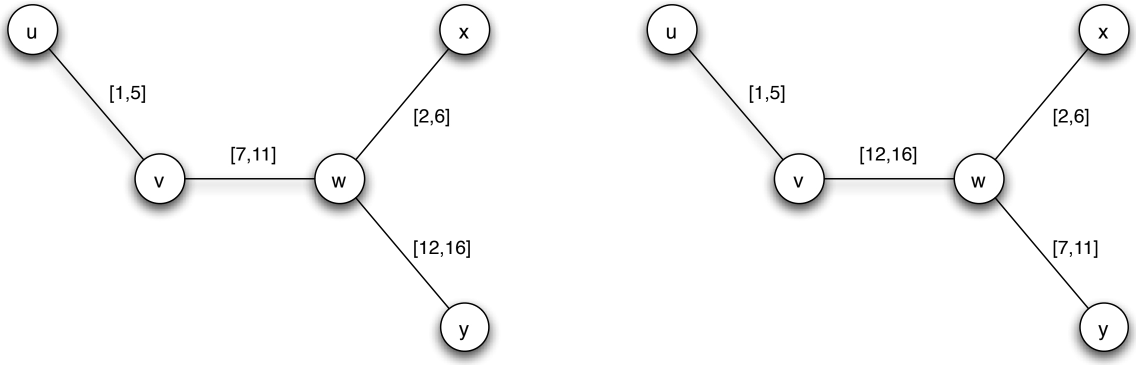
These differences are consistent, however, with the fact that syphilis confers limited temporary immunity after infection, while gonorrhea does not. Moreover, the timing of the syphilis cycles fit well with the timing of the immune properties associated with it. And from the cyclic patterns, one finds that the extent of synchronization between different regions of the United States increases over time, suggesting that the contact network on which it spread became increasingly connected with cross-country links over the second half of the 20th century [195].

There are many further directions in which research on epidemic synchronization is proceeding, including attempts to model more complex temporal phenomena. For example, data for some diseases such as measles shows that epidemics in different cities can synchronize so as to be *out of phase*, with the flare-ups in one city consistently coinciding with troughs in the other [196]. One needs more than simply long-range contacts to explain such properties [213]. There is also the question of how immunization, prevention programs, and other medical interventions can take advantage of these timing properties — another way in which insights from even simple models can help to inform decision-making in this area.

21.6 Transient Contacts and the Dangers of Concurrency

Thus far, our epidemic models have taken the underlying contact network to be a relatively static object, in which all the links in the contact network are present throughout the course of the epidemic. This is a reasonable simplifying assumption for diseases that are relatively contagious and spread quickly, at a rate faster than the typical creation or dissolution of a contact.

But as we move down the spectrum toward diseases that spread through a population over longer time scales, it is useful to revisit these assumptions. For a disease like HIV/AIDS, the epidemic progresses over many years, and its course is heavily dependent on the properties of the sexual contact network. Most people have zero, one, or very few contacts at any single



(a) *In a contact network, we can annotate the edges with time windows during which they existed.*

(b) *The same network as in (a), except that the timing of the w-v and w-y partnerships have been reversed.*

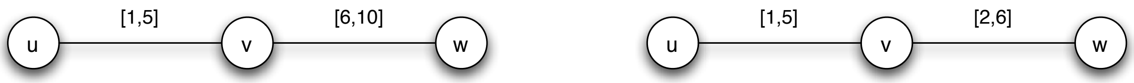
Figure 21.8: Different timings for the edges in a contact network can affect the potential for a disease to spread among individuals. For example, in (a) the disease can potentially pass all the way from u to y, while in (b) it cannot.

point in time (a few people have many, which is important as well); and the identities of these contacts can shift significantly while the disease progresses, as new sexual partnerships are formed and others break up.

So for modeling the contact network in such diseases, it is important to take into account the fact that contacts are *transient* — they do not necessarily last through the whole course of the epidemic, but only for particular windows of time. Thus, we will consider contact networks in which each edge is annotated with the *period of time* during which it existed — that is, the time range over which it was possible for one endpoint of the edge to have passed the disease directly to the other.

Figure 21.8(a) shows an example of this, with the numbers inside square brackets indicating the time ranges when each edge exists. Thus the u-v and w-x partnerships happen first, and they overlap in time; after this, w has a partnership with v and then later with y. Note also that for this section — in keeping with the motivation from HIV/AIDS and similar diseases — we assume the edges to be undirected rather than directed, to indicate that infection can pass in either direction between a pair of people in a partnership. (As in previous sections, we could also accomplish this by having directed edges pointing in both directions between each pair of connected people, but since everything here will be symmetric, it is more convenient to use undirected edges.)

The Consequences of Transient Contacts. A little experimentation with the example in Figure 21.8(a) indicates how the timing of different edges can affect the spread of a disease.



(a) *v's two partnerships happen serially*

(b) *v's two partnership's happen concurrently*

Figure 21.9: A disease tends to be able to spread more widely with concurrent partnerships (b) than with serial partnerships (a).

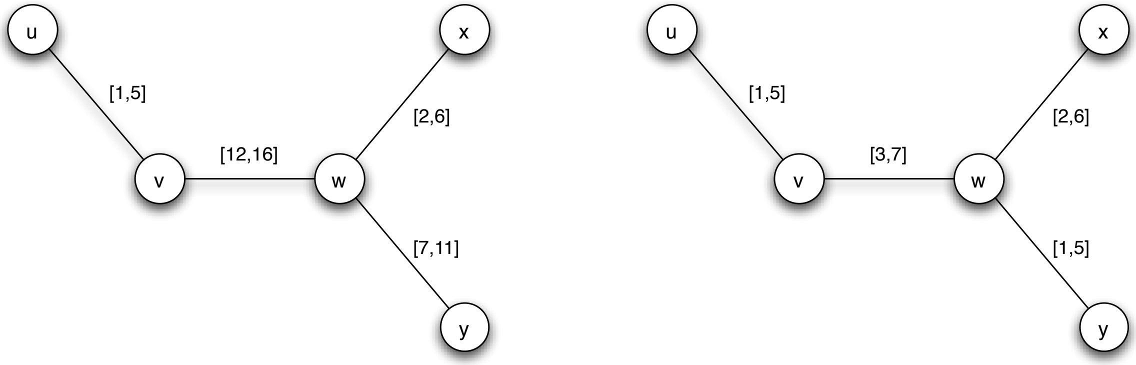
For example, if u has the disease at time 1, it is possible for it to spread all the way to y, through v and w as intermediaries. (Of course, if contagion is probabilistic as before, it will not necessarily succeed in spreading; but it has the potential to do so.) On the other hand, u cannot spread the disease to x: node u could pass the disease to v, who could pass it to w; but by the time it reaches w, the partnership of w and x is long over.

Moreover, changing the timing of partnerships can change the possible transmission pathways, even as the set of underlying contacts remains the same. For instance, the example in Figure 21.8(b) differs from the one in Figure 21.8(a) only in that the temporal order of the w-v and w-y partnerships has been reversed. But notice that while u was able to pass the disease all the way to y in Figure 21.8(a), it cannot do so in Figure 21.8(b): in the latter case, the w-y partnership is over by the time the disease could possibly get from u to w.

Such considerations are crucial as health workers and epidemiologists map out the contact networks associated with a disease such as HIV/AIDS. For example, we can see from the difference between Figures 21.8(a) and 21.8(b) that in order for y to know whether he or she is at risk from a disease carried by u, it is not enough even to map out the full set of sexual partnerships; it is crucial to know information about the order of events as well. Or if we go back to the striking Figure 2.7 from Chapter 2, mapping out the relationships within a high school, we can appreciate that the image itself is not enough to fully chart the potential spread of diseases through this population — we would also need to know the timing of these relationships.

Networks in which the edges only exist for specific periods of time have been the subject of modeling efforts in many areas, including sociology [182, 305, 258], epidemiology [307, 406], mathematics [106], and computer science [53, 239]. It is an issue that is relevant not just to the spread of disease, but also to a wide range of settings that are modeled by networks. For example, the diffusion of information, ideas, and behaviors through social networks clearly also depends on how the timing of different communications between people either enables or blocks the flow of information to different parts of the population.

Concurrency. Differences in the timing of contacts do not just affect who has the potential to spread a disease to whom; the pattern of timing can influence the severity of the overall



(a) *No node is involved in any concurrent partnerships*

(b) *All partnerships overlap in time*

Figure 21.10: In larger networks, the effects of concurrency on disease spreading can become particularly pronounced.

epidemic itself. A timing pattern of particular interest — and concern — to HIV researchers is *concurrency* [307, 406].

A person is involved in *concurrent* partnerships if he or she has two or more active partnerships that overlap in time. For example, in each of Figures 21.9(a) and 21.9(b), node v has partnerships with each of u and w. But in the first of these figures, the partnerships happen serially — first one, then the other — while in the second, they happen concurrently, overlapping in time. The concurrent pattern causes the disease to circulate more vigorously through this three-person network. u and w may not be aware of each other's existence, but the concurrent partnerships make it possible for either of u or w to spread the disease to the other; the serial partnerships only allow spreading from u to w, but not the other way. In larger examples one can find more extreme effects; for example, Figure 21.10(b) differs from Figure 21.10(a) only in that the time windows of the partnerships have been "pushed together" so that they all overlap. But the effect is considerable: where the pattern in Figure 21.10(a) allowed different parts of the network to be "walled off" from each other by the timing effects, the concurrent partnerships make it possible for any node with the disease to potentially spread it to any other.

In simulations with various notions of concurrency, Morris and Kretzschmar found that small changes in the amount of concurrency — keeping other variables like the average number and duration of partnerships fixed — could produce large changes in the size of the epidemic [307]. Qualitatively, this aligns well with the intuition from earlier sections, that changing the average number of new cases of a disease caused by an infected individual even slightly can sometimes have significant consequences. For some of the simplest models, such as the branching process, it is possible to make this intuition precise; for more complex models such as the present one involving concurrency in arbitrary networks, it remains the topic of ongoing research.

Concurrency is just one particular kind of pattern to be found in the timing of relationships in a contact network. Further research in this area could possibly uncover more subtle patterns as well; the interaction of timing and network structure has the potential to provide further insights into the way diseases spread through the changing contacts within a population.

21.7 Genealogy, Genetic Inheritance, and Mitochondrial Eve

Our discussion of epidemics has provided us with a way of thinking about processes that spread randomly over time through a network. As mentioned earlier, this is a useful framework for modeling many kinds of things that spread, not just diseases. The spread of information can be modeled this way, as an alternative to the approaches based on explicit decision rules discussed in Chapter 19. In such settings, adapting the ideas from this chapter can be relatively straightforward but still very informative.

In this section, instead, we apply the perspective of random spreading to a situation where the connection is at first a bit more subtle; it takes a little work to precisely identify the network and the process that is spreading through it. The setting is that of genetic inheritance. What we will find is that when we view inheritance of traits as a random process that takes place on a network linking organisms in successive generations — in other words, with edges connecting parents to their offspring — then we can obtain insight into some fundamental hereditary processes. We start with a story that illustrates some of the basic genetic issues we'll consider.

Mitochondrial Eve. In 1987, Rebecca Cann, Mark Stoneking, and Allan Wilson published a paper in the journal *Nature* [94] where they provided evidence for a rather striking proposition. Consider following your maternal ancestry backward in time through human history, producing a trail that goes from you to your mother, to her mother (i.e. your maternal grandmother), to *her* mother, and so on indefinitely. Each of us in principle can produce such a maternal ancestry trail, which we'll call a maternal *lineage*. Now, the claim of Cann, Stoneking, and Wilson was that all these lineages in fact meet at a single woman who lived between 100,000 and 200,000 years ago, probably in Africa. She is at the root of all our maternal ancestries.

Let's first ask how they reached this conclusion, and then consider what it signifies. One way to infer facts about maternal ancestries is to study the DNA found not in our cells' nuclei, but in the much smaller, separate genome that each of us has in our cells' mitochondria. Unlike nuclear DNA, which contains parts of both our parents' genomes, this mitochondrial DNA is (to a first approximation) passed to children entirely from their mothers. So roughly speaking, aside from random mutations, you have your mother's DNA, she has her mother's DNA, and so on through your maternal ancestry. With this in mind, Cann, Stoneking, and Wilson analyzed the mitochondrial DNA of people drawn from a wide sample of geographic and ethnic backgrounds; using standard techniques to estimate the rate at which genetic sequences will diverge through random mutations over many generations, they concluded that all the mitochondrial DNA in this population likely had a common origin roughly 100,000-200,000 years ago. By "common origin" here, we mean a single mitochondrial genome belonging to a single human being; because she is the source of the mitochondrial DNA of everyone on earth, researchers standardly refer to this woman as *Mitochondrial Eve*.

This finding caught the public imagination when it was first announced; it received a fair amount of media attention at the time, and its implications have been nicely explored in general books about human history [333]. The analysis involved in the original finding has since been refined by a number of other research groups; caveats have been introduced due to the fact that the inheritance of mitochondrial DNA may be more complicated than originally thought; but the basic conclusion has been mainly accepted at a general level.

As to what this finding signifies: on first hearing, it takes a bit of thought to sort out what it implies and what it doesn't. It is indeed striking to be able to posit the existence of a single person from the not-so-distant evolutionary past who is an ancestor of everyone. Mitochondrial Eve (in contrast to her namesake Eve from the Bible) was not asserted to be the only living woman in her time; there were presumably many other women living at the same time as her, but *from the point of view of present-day mitochondrial DNA, all these women are genetically irrelevant:* somewhere along the line from then to now, each of their lines of mitchondrial DNA died out.

On the other hand, one also needs to be careful before attributing too much to the relatively recent existence of Mitochondrial Eve. In particular, while her contemporaries were genetically irrelevant to our mitochondrial DNA, they are not irrelevant to the remainder of our genomes; each of us has genetic contributions from a large number of ancestors. (Though even here there is more going on than meets the eye, as we'll discuss shortly.) Moreover, the overlapping patterns of our respective ancestries are complex and still not well understood; what we learn from Mitochondrial Eve is that all our ancestries are pinned together along their maternal lines, a couple of hundred thousand years into the past.

Ultimately, the identification of Mitochondrial Eve was in a sense a showcase of ideas that had been emerging in the genetics community over the previous decade [245, 325]. These ideas were based on models that could predict the existence of common ancestors and make estimates about their recency. And they showed that at a mathematical level,

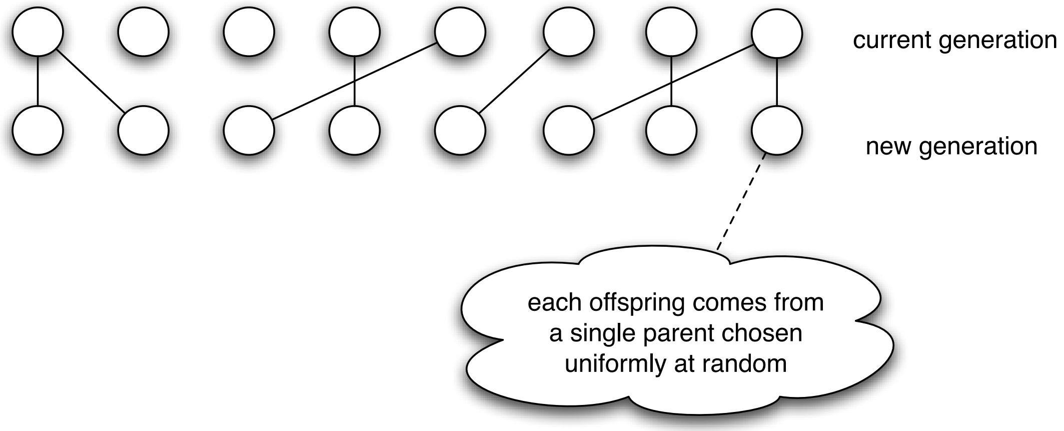

current generation
new generation
each offspring comes from a single parent chosen uniformly at random

Figure 21.11: In the basic Wright-Fisher model of single-parent ancestry, time moves step-by-step in generations; there are a fixed number of individuals in each generation; and each offspring in a new generation comes from a single parent in the current generation.

independent of the difficulty of establishing evidence from genetic data, the existence of someone like Mitochondrial Eve was not only natural, but in fact — as we will see next — essentially inevitable. At their core, these models were built from a probabilistic formalism involving networks; indeed, even in a qualitative sense, one can appreciate something epidemic-like about the way in which copies of different people's mitochondrial DNA spread through subsequent generations, inhabiting future offspring, until one eventually crowds out all the others. We now describe the basic versions of these models, and how they connect to questions about ancestry.

A Model of Single-Parent Ancestry We use a fundamental model of ancestry known in population genetics as the *Wright-Fisher model* [325]. To remain tractable, the model involves a number of simplifying assumptions. Consider a population that is constrained by resources to maintain a fixed size N in each generation. Time moves step-by-step from one generation to the next; each new generation is formed by having the current set of N individuals produce N offspring in total. Each offspring in this new generation is produced from a single parent, and this parent is selected independently and uniformly at random from among those in the current generation. Figure 21.11 depicts this process; as shown there, we can draw the relationship of one generation to the next as a graph, with a node for each individual, and an edge connecting each offspring to their parent chosen uniformly at random from the previous generation. Notice that because of this rule for selecting parents, certain individuals in the upper generation can have multiple children (such as the first and last in Figure 21.11), while others may have none.

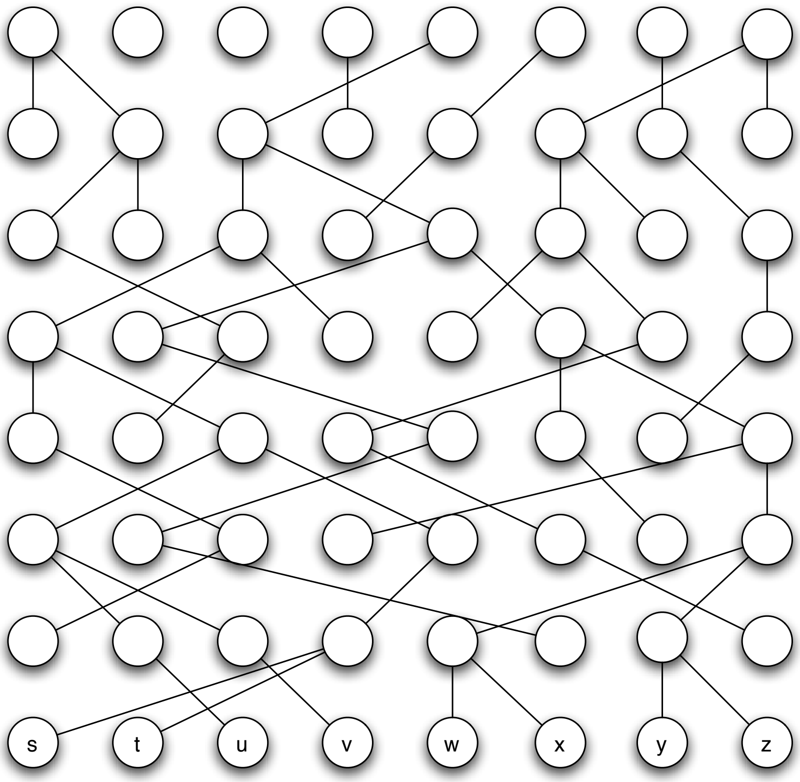


Figure 21.12: We can run the model forward in time through a sequence of generations, ending with a set of present-day individuals. Each present-day individual can then follow its single-parent lineage by following edges leading upward through the network.

The structure of this model reflects a few underlying assumptions. To begin with, we're assuming a *neutral model* in which no individual has a selective advantage in reproduction; everyone has the same chance of producing offspring. Furthermore, we're modeling a situation in which each individual is produced from a *single* parent, as opposed to two parents in a sexually reproducing population. This is consistent with several possible interpretations.

- First, and most directly, it can be used to model species that engage in asexual reproduction, with each organism arising from a single parent.

- Second, it can be used to model single-parent inheritance even in sexually reproducing populations, including the inheritance of mitochondrial DNA among women as in our discussion above. In this interpretation, each node represents a human woman, with women linked to their mothers in the previous generation. Moreover, as we will discuss later, there is in fact a much more general way to use this model to think about inheritance in sexually reproducing populations.

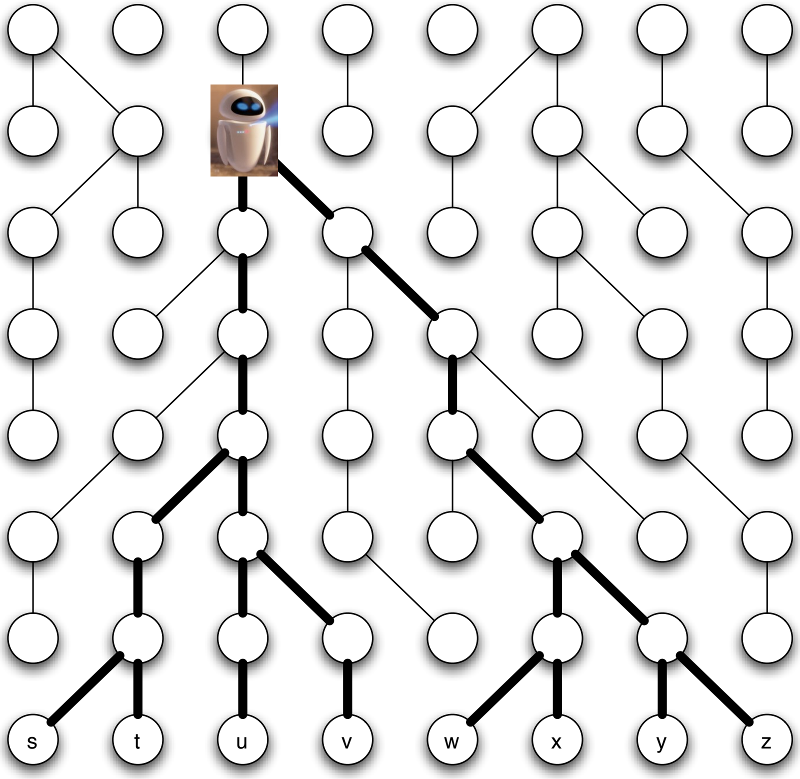


Figure 21.13: A re-drawing of the single-parent network fom Figure 21.12. As we move back in time, lineages of different present-day individuals coalesce until they have all converged at the most recent common ancestor.

- Third, it can be used to model purely "social" forms of inheritance, such as master-apprentice relationships. For example, if you receive a Ph.D. in an academic field, you generally have a single primary advisor. If you model students as being "descended" from advisors, than we can trace ancestries through sequences of advisors back into the past — just as we traced maternal lineages.

Now, if we run this model forward in time through multiple generations, we get a network such as the one pictured in Figure 21.12. Each individual is connected to one parent in the previous generation; time runs from top to bottom, with N present-day individuals in the lowest layer (named s through z in the figure). Notice that from any one of these individuals at the bottom, we can trace its single-parent lineage backward in time by following edges upward, always taking the single edge leading up out of each node we encounter.

If we imagine the individuals in the bottom row of Figure 21.12 to be present-day women, then Mitochondrial Eve would be the lowest node in the figure where all the maternal lineages first fully converge. It's a bit tricky, visually, to find this node in Figure 21.12, but we can re-draw the same ancestries with the paths unscrambled in Figure 21.13, and then the location of Eve becomes easy to see: she's the third node in the second row from the top (with the lineages leading back to her consisting of the darkened edges).

These examples indicate how the existence of common ancestors — and even the number of generations needed to reach them — can be predicted from the Wright-Fisher model. To do this, we employ a useful trick for reasoning about the model: we think of the ancestries as being built *backward* in time, rather than forward. In other words, an equivalent view of the model is to take a set of present-day individuals, and construct earlier generations one at a time by having each current individual choose its parent in the previous generation independently and uniformly at random.

We can see how this works by moving upward through the levels of Figure 21.13. Whenever two individuals happen to choose the same parent, then their lineages *coalesce* into a common lineage from that point onward. Thus, we start with N distinct lineages in the present, but as we build generations going backward in time, the number of distinct lineages decreases whenever individuals on distinct lineages choose the same parent. This coalescence will happen rapidly at first, when there are many lineages and the probability of a collision between lineages is high; as time goes on, the number of distinct lineages of present-day individuals shrinks more and more slowly. But whenever there is more than one distinct lineage, there is a finite expected time until two of them collide, and so the process must eventually reach a single lineage. The node at which this first happens is called the *most recent common ancestor* — the analogue of Mitochondrial Eve in this model. The model is simple enough that one can estimate the expected time until the collisions among lineages, and hence the expected number of generations to the most recent common ancestor [245, 325].

Genetic Interpretations. Although the maternal inheritance of mitochondrial DNA makes for a very simple single-parent process, the Wright-Fisher model is relevant to sexually reproducing populations for a much more fundamental reason. While the chromosomes of your parents recombined to produce your genome, making your chromosomes a patchwork of theirs, any *single* point in your genome — a single nucleotide on one of your chromosomes — was inherited from just one of your mother or father. They, in turn, inherited it from just one of their mother or father, and so on. As a result, if we want to trace the ancestry of a single point in your genome, we are following a single-parent lineage, even though offspring are produced by sexual reproduction. The most recent common ancestor for this particular point, looking across a population of N individuals, will thus follow from the same analysis we've seen above, as it did for mitochrondrial DNA.

Because of recombination, the lineages for one point in the genome may differ from the lineages for even a nearby point, and hence the most recent common ancestors may differ as well. One can develop probabilistic models for how these lineages relate to each other, but the analysis becomes much more complex [418].

There are many other issues that arise when extending these simplified models to more complex genetic applications. For example, geographic barriers in a population can isolate individuals from each other, and this can have an effect on the patterns of interaction among lineages [354]. More generally, spatial constraints on the interactions among individuals can affect these patterns, providing another setting in which network properties can potentially inform broader conclusions about genetic outcomes.

21.8 Advanced Material: Analysis of Branching and Coalescent Processes

In this section, we analyze two of the basic processes discussed in this chapter: the *branching process* for the spread of an epidemic with simplified contact network structure, and the *coalescent process* for the merging of lineages back to a common ancestor. Both of these are based on probabilistic reasoning involving branching tree structures: the first as the epidemic spreads forward through individuals, and the second as the lineages travel backward in time.

A. Analysis of Branching Processes

Recall the branching process model that we considered in Section 21.2: each infected individual meets k others and infects each with probability p. Thus, the expected number of new cases of the disease caused by each infected individual is $R_0 = pk$, the *basic reproductive number*. We want to show that the persistence of the disease depends critically on whether R_0 is smaller or larger than 1, a notion that we will formulate as follows.

Recall that the population in this model is organized into a *tree* (as shown in Figure 21.1(a)) in which every node is connected to k nodes just below it. Let q_n denote the probability that the epidemic survives for at least n waves — in other words, that some individual in the n^{th} level of the tree becomes infected. Let q^* be the limit of q_n as n goes to infinity; we can think of this as the probability that the disease persists indefinitely. We will prove the following claim.

Claim: (a) If $R_0 < 1$ then $q^* = 0$. (b) If $R_0 > 1$ then $q^* > 0$.

This establishes the “knife-edge” quality of R_0 that we discussed in Section 21.2.

The Expected Number of Infected Individuals. We start by considering an approach to this problem that gets us partway to a proof of the claim: considering the expected number of infected individuals at each level of the tree.

First, let’s consider the total number of individuals at each level. The number of individuals at any given level exceeds the number at the previous level by a factor of k, and

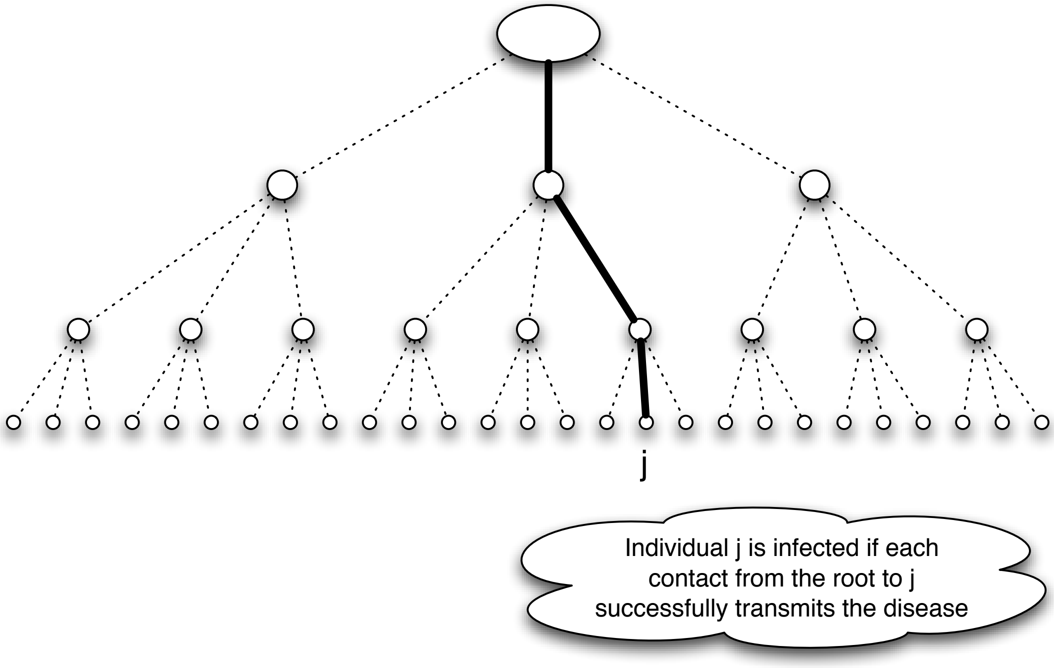

j
Individual j is infected if each contact from the root to j successfully transmits the disease

Figure 21.14: To determine the probability that a particular node is infected, we multiply the (independent) probabilities of infection on each edge leading from the root to the node.

therefore the number who are at level n is k^n. (This is also true at level $n = 0$: the top level consists of just the root, and $k^0 = 1$.)

Now, let X_n be a random variable equal to the number of infected individuals at level n. One way to think about the expected value $E[X_n]$ is to write X_n as a sum of simpler random variables as follows. For each individual j at level n, let Y_{nj} be a random variable equal to 1 if j is infected, and equal to 0 otherwise. Then

$$X_n = Y_{n1} + Y_{n2} + \cdots + Y_{nm},$$

where $m = k^n$, since the right-hand side simply counts up, one-by-one, the number of infected individuals at level n. Linearity of expectation says that the expectation of the sum of a set of random variables is equal to the sum of their expectations, and so

$$E[X_n] = E[Y_{n1} + Y_{n2} + \cdots + Y_{nm}] = E[Y_{n1}] + E[Y_{n2}] + \cdots + E[Y_{nm}]. \qquad (21.1)$$

The reason to write things this way is that each expectation on the right-hand side is extremely easy to work out: $E[Y_{nj}] = 1 \cdot \Pr[Y_{nj} = 1] + 0 \cdot \Pr[Y_{nj} = 0] = \Pr[Y_{nj} = 1]$, and so the expectation of each Y_{nj} is just the probability that individual j gets infected.

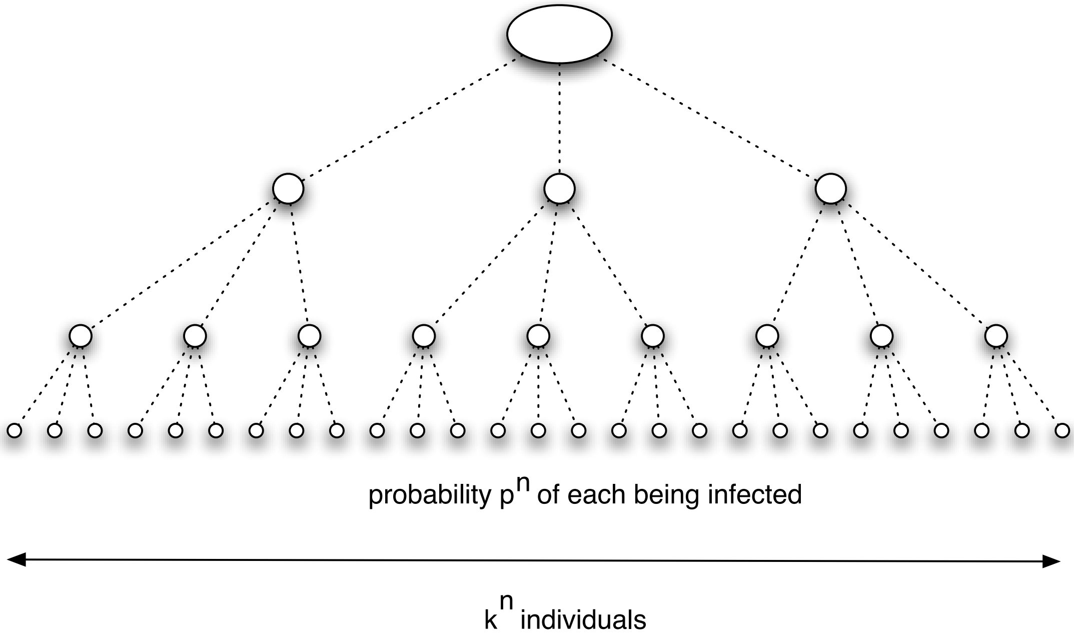


Figure 21.15: The expected number of individuals infected at level n is the product of the number of individuals at that level (k^n) and the probability that each is infected (p^n).

Individual j at depth n gets infected precisely when each of the n contacts leading from the root to j successfully transmit the disease, as shown in Figure 21.14. Since each contact transmits the disease independently with probability p, individual j is infected with probability p^n. Therefore $E[Y_{nj}] = p^n$. We have already concluded that there are k^n individuals at level n of the tree, and hence k^n terms on the right-hand side of Equation (21.1). Therefore, as summed up in Figure 21.15, we conclude that

$$E[X_n] = p^n k^n = (pk)^n = R_0^n. \tag{21.2}$$

From Expected Values to Probabilities of Persistence. Equation (21.2) suggests the importance of the basic reproductive number R_0 in reasoning about the spread of an epidemic in the branching process model. Now let's consider what this tells us about q^*, the probability that the epidemic persists indefinitely.

First, the fact that $E[X_n] = R_0^n$ immediately establishes part (a) of the Claim, that when $R_0 < 1$ we have $q^* = 0$. To see why, we go back to the definition of $E[X_n]$ and apply a fact that we also found useful in Section 20.7. To recap the discussion there, the definition for the expected value is

$$E[X_n] = 1 \cdot \Pr[X_n = 1] + 2 \cdot \Pr[X_n = 2] + 3 \cdot \Pr[X_n = 3] + \cdots \tag{21.3}$$

and an alternate but equivalent way to write the right-hand side is as

$$\Pr[X_n \geq 1] + \Pr[X_n \geq 2] + \Pr[X_n \geq 3] + \cdots \tag{21.4}$$

since we observe that each term $\Pr[X_n = i]$ contributes exactly i copies of itself to the sum in (21.4). Therefore we have

$$E[X_n] = \Pr[X_n \geq 1] + \Pr[X_n \geq 2] + \Pr[X_n \geq 3] + \cdots \tag{21.5}$$

From (21.5) we observe that $E[X_n]$ must be at least as large as the first term on the right-hand side, and so $E[X_n] \geq \Pr[X_n \geq 1]$. Notice also that $\Pr[X_n \geq 1]$ is precisely the definition of q_n, and so $E[X_n] \geq q_n$. But $E[X_n] = R_0^n$ which is converging to 0 as n grows, and hence q_n must also be converging to 0. This shows that $q^* = 0$ when $R_0 < 1$.

Now, when $R_0 > 1$, the expected values $E[X_n] = R_0^n$ go to infinity as n grows. *However*, this fact by itself is not enough to show that $q^* > 0$. It is entirely possible to have a sequence of random variables for which $E[X_n]$ goes to infinity but $\Pr[X_n > 0]$ converges to 0 as n grows. (As a simple example, suppose that X_n were a random variable taking the value 4^n with probability 2^{-n}, and taking the value 0 otherwise. Then $E[X_n] = (4/2)^n = 2^n$, which goes to infinity, while $\Pr[X_n > 0] = 2^{-n}$, which goes to 0.)

This won't happen in our case, but these considerations do say that to establish $q^* > 0$ when $R_0 > 1$, we'll need to use something more specific about the process than simply the expected number of infected individuals. We do this now, developing a formula for q_n that in the end will allow us to determine the value of q^* exactly.

A Formula for q_n. The quantity q_n depends on three more fundamental quantities: the number of contacts per individual k, the contagion probability p, and the level of the tree n. In fact, it's difficult to write down a direct formula for q_n in terms of these quantities, but it's not hard to express q_n in terms of q_{n-1}. This is what we'll do first.

Consider the root node, and let's first ask what it would take for the following event to hold:

> ($*$) The disease spreads through the root node's first contact j and then continues to persist down to n levels *in the part of the tree reachable through j.*

This is illustrated in Figure 21.16. First, for the event ($*$) to hold, it would require that j catches the disease directly from the root, which happens with probability p. At this point, j becomes completely analogous to the root node of its own branching process, consisting of all nodes reachable from it downward in the tree. So for event ($*$) to hold, after j is infected, it is then necessary that the disease persists for $n-1$ levels in the version of the branching process in which we view node j as the root. This happens with probability q_{n-1},

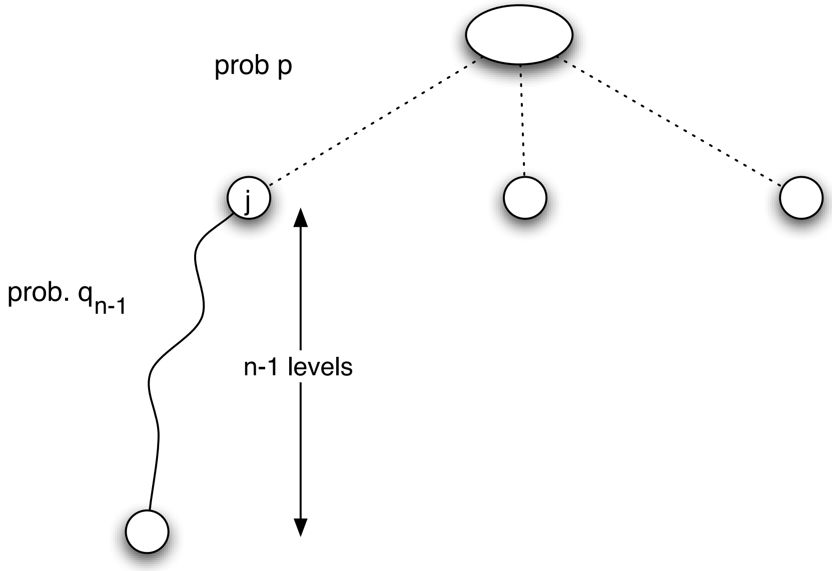


Figure 21.16: In order for there to be an infection at level n, the root must infect one of its immediate descendants, and then this descendant must, recursively, produce an infection at level $n-1$.

by the definition of q_{n-1}. Therefore, the probability of the event $(*)$ is pq_{n-1}. Or, taking the complementary view, event $(*)$ fails to hold with probability

$$1 - pq_{n-1}.$$

Now, there is a copy of event $(*)$ for each of the direct contacts of the root node, and each fails to hold with probability $1-pq_{n-1}$. Since they're independent, the probability that they *all* fail to hold is

$$(1 - pq_{n-1})^k.$$

At this point, we're almost done. The disease fails to persist down to level n of the tree, starting at the root, if it fails to reach level n through any of the root's direct contacts. In other words, the disease fails to persist to level n precisely when all the copies of event $(*)$, for each direct contact of the root, fail to hold. Again, we just determined that this probability is $(1 - pq_{n-1})^k$. But this probability is also $1 - q_n$, since by the definition of q_n, the quantity $1 - q_n$ is exactly the probability that the disease fails to persist to n levels. Therefore,

$$1 - q_n = (1 - pq_{n-1})^k$$

and solving for q_n we get

$$q_n = 1 - (1 - pq_{n-1})^k. \tag{21.6}$$

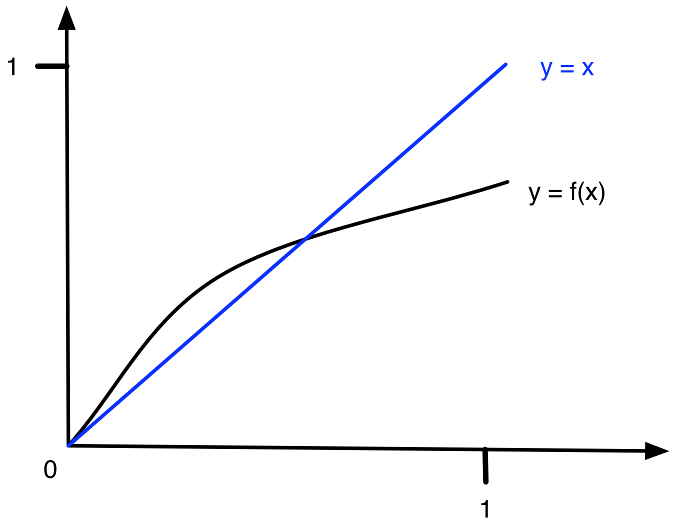


Figure 21.17: To determine the limiting probability of an infection at depth n, as n goes to infinity, we need to repeatedly apply the function $f(x) = 1 - (1 - px)^k$, which is the basis for the recurrence $q_n = f(q_{n-1})$.

Since we are assuming that the root is infected, and we can treat the root as level 0 of the tree, we have $q_0 = 1$; this simply says that the root is infected with probability 1. Starting from $q_0 = 1$, we can then build up the values $q_1, q_2, q_3, \ldots$ in order, determining each from the previous one in the list using Equation (21.6). Simply being able to determine the values of each q_n this way, however, doesn't immediately tell us where they're headed as n goes to infinity. For that we need a technique that looks at the limiting value of this sequence.

Following the values q_n to a limit. If we define the function $f(x) = 1-(1-px)^k$, then we can write Equation (21.6) as follows: $q_n = f(q_{n-1})$. This suggests a very clean, purely algebraic way of formulating our question about q^*. We have a function $f(x) = 1-(1-px)^k$, and we simply want to study the sequence of values $1, f(1), f(f(1)), f(f(f(1))), \ldots$, obtained by applying f repeatedly.

To get started thinking about this, let's plot the function f on a pair of x-y axes, as in Figure 21.17. Here are some basic facts about f that help in producing this plot.

- First, $f(0) = 0$ and $f(1) = 1 - (1 - p)^k < 1$. This means that the plot of f passes through the origin, but lies below the line $y = x$ once $x = 1$, as shown in Figure 21.17.

- Second, the derivative of f is $f'(x) = pk(1 - px)^k$. Notice that as x ranges between 0 and 1, the quantity $f'(x)$ is positive but monotonically decreasing. This means that f has the increasing but concave shape depicted in Figure 21.17.

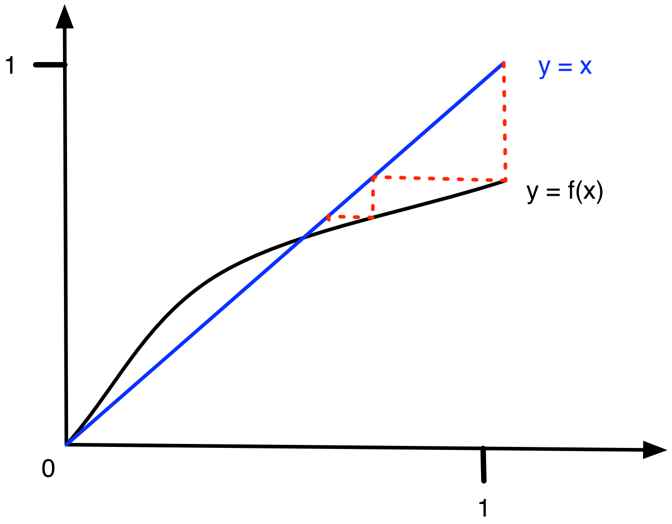


Figure 21.18: When we repeatedly apply the function $f(x)$, starting at $x = 1$, we can follow its trajectory by tracing out the sequence of steps between the curves $y = f(x)$ and $y = x$.

- Finally, the slope of f at $x = 0$ is equal to $f'(0) = pk = R_0$. So in the case when $R_0 > 1$, which is what we're focusing on now, the function f starts out above the line $y = x$ for small positive values of x.

When $R_0 > 1$, we can take these points together — that $y = f(x)$ starts out above $y = x$ for small positive values of x but ends up below it by the time we get to $x = 1$ — and conclude that $y = f(x)$ must cross $y = x$ somewhere in the interval between 0 and 1, at a point $x^* > 0$.

Now, using this plot, let's take a geometric view of the sequence of values

$$1, f(1), f(f(1)), f(f(f(1))), \ldots$$

that we're analyzing. In particular, let's track this sequence on the line $y = x$. If we're currently at a particular point (x, x) on the line $y = x$, and we want to get to the point $(f(x), f(x))$, we can do that as follows. We first move vertically to the curve $y = f(x)$; this puts us at the point $(x, f(x))$. We then move horizontally back to the line $y = x$; this puts us at the point $(f(x), f(x))$ as desired. This two-step vertical-horizontal motion is depicted as the first two parts of the dashed line in Figure 21.18. Continuing this process, we pass through all the points in the sequence $x, f(x), f(f(x)), \ldots$ along the line $y = x$.

If we start this from $x = 1$, as indicated in Figure 21.18, the process converges to the point (x^*, x^*) where the line $y = x$ meets the curve $y = f(x)$. Now we can go back to the interpretation of all this in terms of the branching process. The sequence of values

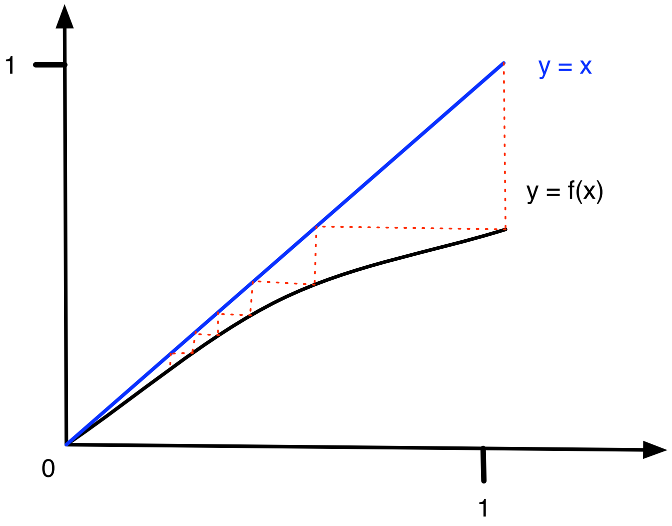


Figure 21.19: When $y = f(x)$ only intersects $y = x$ at zero, the repeated application of $f(x)$ starting at $x = 1$ converges to 0.

$1, f(1), f(f(1)), \ldots$ is precisely the sequence $q_0, q_1, q_2, \ldots$, as we argued above, and so we have concluded that it converges to $x^* > 0$: the unique point at which $f(x) = x$ in the interval strictly between 0 and 1.

This concludes the argument that when $R_0 > 1$, the probability that the epidemic persists for n levels converges to a positive value as n goes to infinity.

It is also worth noticing that this style of analysis shows that $q^* = 0$ when $R_0 < 1$. Indeed, when $R_0 < 1$, the curve $y = f(x)$ looks much like it does in Figure 21.17, *except* that its derivative at 0 is $R_0 < 1$, and so it lies out below the line $y = x$ for the whole interval between 0 and 1. This means that when we follow the sequence of values $1, f(1), f(f(1)), \ldots$ as the dashed lines do in Figure 21.18, it descends all the way to $x = 0$ without stopping at any intermediate point. (See Figure 21.19.) This shows that the resulting limit, which is q^*, is equal to 0 in this case.

B. Analysis of Coalescent Processes

We now analyze a different process arising from earlier in the chapter — the merging of ancestral lineages discussed in Section 21.7. In particular, we will derive an estimate of the expected number of generations one has to go back in order to find the most recent common ancestor for a set of individuals in the model from that section [245, 325]. Like the analysis of branching processes, this will require probabilistic calculations on trees. In this case, however, it is tricky to get an exact answer, and we will make use of two approximations as

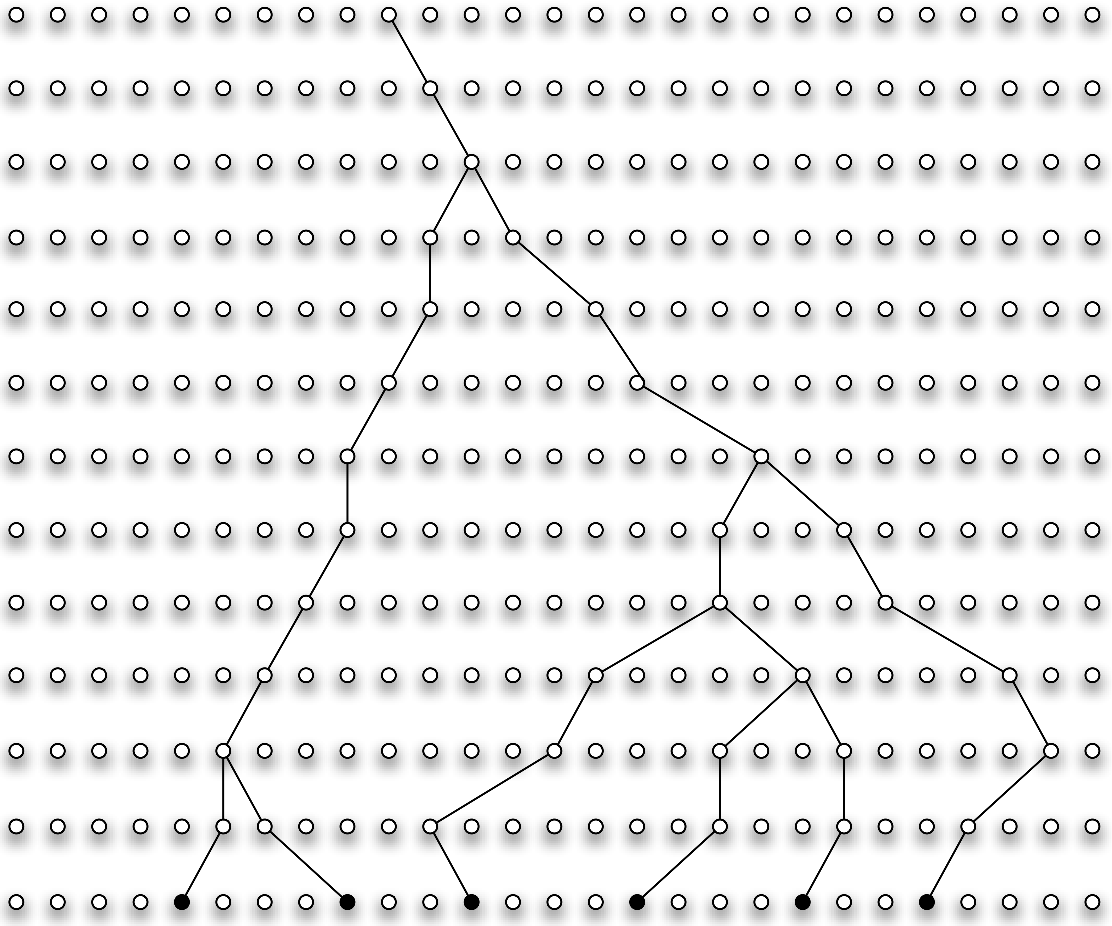


Figure 21.20: We can view the search for coalescence as a backward walk through a sequence of earlier generations, following lineages as they collide with each other.

we estimate the required number of generations. (In fact, these approximations still allow for a very accurate estimate.)

In addition to the approximations, which we'll specify in context later, we start by varying the statement of the problem slightly, following the original work on the topic. Specifically, we will focus on a small sample of k individuals in a large population of size N; rather than analyzing the time until all lineages in the full population merge into a common ancestor, we will consider the time until the lineages of these k merge into a common ancestor. This is reasonable from the point of view of applications, since generally one is only ever studying a fixed-size sample of a large population; also, the calculations involved provide insight into the question for the full population as well.

To recall the model from Section 21.7, adapted to the plan of looking at fixed-size samples of k individuals, we can pose the question as follows. There are N individuals in each generation. For each of the k individuals in the initial sample, we choose a parent for each uniformly at random from the previous generation. We continue working backward in time this way, extending each of the k lineages through earlier generations. Whenever we get to a generation where two individuals happen to choose the same parent, their lineages merge (since their ancestors will now be the same), and so the process continues with fewer distinct lineages to track. Finally, we stop when we first reach a point where the number of lineages has been reduced to one, a moment that we call *coalescence*. We want to estimate the expected time until coalescence occurs. Figure 21.20 illustrates this on an example with $k = 6$ initial present-day individuals (in the bottom row) and a population size of $N = 27$ (the number of nodes in each row).

The Probability that Lineages Collide in One Step. The key to this analysis is to consider a single step, in which we have a set of j distinct lineages that we're tracking, and we want to estimate the probability that at least two of them will choose a common parent in the previous generation. (We will call this a *collision* between two lineages.)

The easiest case to think about is $j = 2$. Suppose we view the random choice as being made sequentially by the two lineages under consideration. The first lineage chooses a parent uniformly at random, and then there is only a collision if the second lineage picks the same parent uniformly at random from the N available choices. The probability that this occurs is therefore exactly $1/N$.

Things get more complicated when j is larger than 2. First, let's compute the probability that no two lineages collide by imagining that the lineages choose their parents one at a time. For no two lineages to collide, it must be the case that after the first lineage chooses a parent, the second chooses a parent distinct from this, the third chooses a parent distinct from these two, and so forth, up to the j^{th} lineage, who must choose a parent distinct from the first $j - 1$. The probability that this happens is

$$\left(1 - \frac{1}{N}\right)\left(1 - \frac{2}{N}\right)\left(1 - \frac{3}{N}\right)\cdots\left(1 - \frac{j-1}{N}\right).$$

Expanding out this product, we see that it is equal to

$$1 - \left(\frac{1 + 2 + 3 + \cdots + j - 1}{N}\right) + (\text{terms with } N^2 \text{ or higher in the denominator}).$$

In particular, it is at most

$$1 - \left(\frac{1 + 2 + 3 + \cdots + j - 1}{N}\right) + \frac{g(j)}{N^2}$$

for a function $g(\cdot)$ that depends only on j. So far this calculation has been exact, but here we come to the first of two approximations, following [245]: rather than deal with the complexity of this last term, we observe that when the population size N is much larger than j, expressions of the form $g(j)/N^2$ are negligible compared to $(1+2+\cdots+j-1)/N$. We therefore ignore them, and approximate the probability that no two lineages collide by

$$1-\left(\frac{1+2+3+\cdots+j-1}{N}\right)=1-\frac{j(j-1)}{2N}. \tag{21.7}$$

Now, when two lineages do in fact collide, there are a number of possibilities: it could be that there is simply a two-way collision between two of the lineages while all the others remain distinct, or it could be that more than two lineages collide in a single generation. We now describe how the latter scenario can come to pass, and then argue that it is very unlikely.

- First, it could be that three lineages all choose the same parent in a single generation. For any particular set of three lineages, the probability this happens is exactly $1/N^2$: imagining the choice being made sequentially, the first lineage can pick any parent, and then the second and third must independently pick this same parent from the N available choices. Now, since there are fewer than j^3 sets of three lineages, the probability that any three-way collision happens in a given generation is less than j^3/N^2. When N is much larger than j, this quantity is negligible compared to expressions as in (21.7) that only have N in the denominator.

- Alternately, it could be that two different pairs of lineages each have a separate, two-way collision in the same generation: suppose that lineage A collides with lineage B, and lineage C collides with lineage D. The collision of A and B has probability $1/N$, and the collision of C and D is an event independent from this, also with probability $1/N$. Therefore, for this particular choice of four lineages, both collisions happen with probability $1/N^2$. Since there are less than j^4 ways of choosing A, B, C, and D, the probability that there are simultaneous two-way collisions involving any choice of four lineages is less than j^4/N^2. Again, with N much larger than j, this quantity is negligible compared to expressions that only have N in the denominator.

These arguments lead to our second approximation: we will assume there is never a generation prior to the most recent common ancestor in which we have more than a single two-way collision among lineages.

This means that when we have a generation in which the j current lineages fail to remain completely distinct, it happens because exactly two of them choose a common parent, reducing the number of lineages from j to $j-1$.

The Expected Time Until Coalescence. Our two approximations have led to a very clean view of the process as it works backward in time. In this approximate view, we start

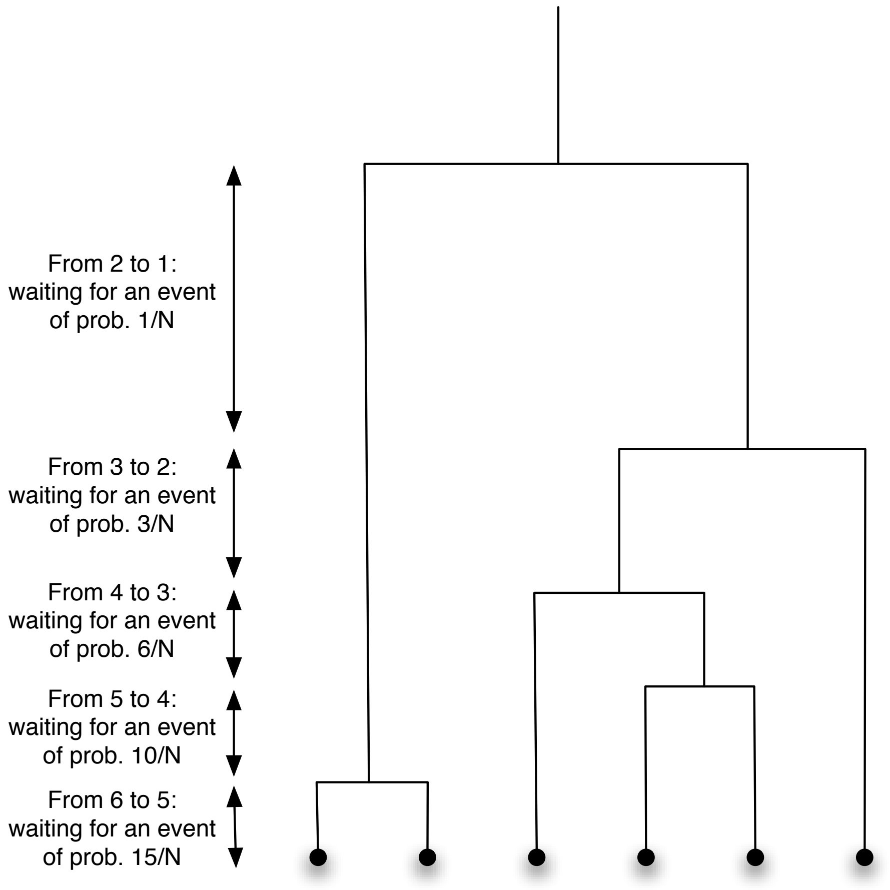


Figure 21.21: Assuming that no three lineages ever collide simultaneously, the time to coalescence can be computed as the time for a sequence of distinct collision events to occur.

with k distinct lineages and wait until two of them collide. This happens with probability $\frac{k(k-1)}{2N}$ in each generation. Once a collision happens, we have $k-1$ distinct lineages, and we wait for two of them to collide with probability $\frac{(k-1)(k-2)}{2N}$ per generation. Things continue this way until we are down to two distinct lineages, at which point we wait for them to collide with probability $\frac{2}{2N} = \frac{1}{N}$ per generation. The overall process is shown, for our example with $k = 6$, in Figure 21.21.

Guided by this view of the process, we can analyze it as follows. Let W be a random variable equal to the number of generations back until coalescence. We can write

$$W = W_k + W_{k-1} + W_{k-2} + \cdots + W_2,$$

where W_j is a random variable equal to the number of generations during which there are exactly j distinct lineages. By linearity of expectation, we have

$$E[W] = E[W_k] + E[W_{k-1}] + \cdots + E[W_2].$$

So it remains to figure out the terms on the right-hand side.

Each random variable of the form W_j can be viewed in the following way: when we have j distinct lineages, we move backward through successive generations, waiting until a particular event (a collision) first happens. We now make use of our approximations: W_j is very close to a simpler random variable in which we have j lineages, we count the number of steps until this number of lineages is reduced to $j-1$, and in each generation this reduction of lineages occurs with probability exactly equal to $p = \frac{j(j-1)}{2N}$. We let X_j denote this closely related, simpler random variable; we write

$$X = X_k + X_{k-1} + X_{k-2} + \cdots + X_2,$$

and we will be interested in determining the expectation

$$E[X] = E[X_k] + E[X_{k-1}] + \cdots + E[X_2]$$

rather than the expectation $E[W]$.

How should we think about the expectation of one of these simpler random variables X_j? It is precisely as though we have a coin that comes up "heads" with a fixed probability $p = \frac{j(j-1)}{2N}$ per flip, and we want to know the expected number of flips until we see the first heads. To compute this expectation, we recall Equation (21.5) from earlier in this section, applied to the current random variable X_j:

$$E[X_j] = \Pr[X_j \geq 1] + \Pr[X_j \geq 2] + \Pr[X_j \geq 3] + \cdots$$

The probability that X_j is at least some value i is just the probability that the coin comes up "tails" on its first i flips, which is $(1-p)^i$. Therefore,

$$E[X_j] = 1 + (1-p) + (1-p)^2 + (1-p)^3 + \cdots = \frac{1}{1-(1-p)} = \frac{1}{p}.$$

This is a very intuitive relationship: the expected time to see the first heads on a coin with a heads probability of p is simply $\frac{1}{p}$.

The random variable X_j describes precisely this process with $p = \frac{j(j-1)}{2N}$. Therefore,

$$E[X_j] = \frac{2N}{j(j-1)},$$

and so

$$\begin{aligned} E[X] &= \frac{2N}{2\cdot 1} + \frac{2N}{3\cdot 2} + \frac{2N}{4\cdot 3} + \cdots + \frac{2N}{j(j-1)} + \cdots + \frac{2N}{k(k-1)} && (21.8) \\ &= 2N\left(\frac{1}{2\cdot 1} + \frac{1}{3\cdot 2} + \frac{1}{4\cdot 3} + \cdots + \frac{1}{j(j-1)} + \cdots + \frac{1}{k(k-1)}\right). && (21.9) \end{aligned}$$

This last sum can be evaluated by noticing that

$$\frac{1}{j(j-1)} = \frac{1}{j-1} - \frac{1}{j}$$

and applying this identity to each term in (21.9) we get

$$E[X] = 2N\left(\left[\frac{1}{1} - \frac{1}{2}\right] + \left[\frac{1}{2} - \frac{1}{3}\right] + \cdots + \left[\frac{1}{j-1} - \frac{1}{j}\right] + \left[\frac{1}{j} - \frac{1}{j+1}\right] + \cdots + \left[\frac{1}{k-1} - \frac{1}{k}\right]\right).$$

In this new way of writing the sum, almost all the terms inside the parentheses cancel each other out, and the only two that survive are 1 and $-\frac{1}{k}$. Hence we conclude that

$$E[X] = 2N\left(1 - \frac{1}{k}\right).$$

This gives us the result we were looking for — the approximate number of generations until coalescence — and so we can conclude with just a few final observations. First, once k becomes moderately large, the expected time to coalescence depends only very weakly on k; it is roughly $2N$ as k grows. Second, the breakdown of X into $X_k + X_{k-1} + \cdots + X_2$ lets us appreciate where most of the time is being spent in the merging to a common ancestor. As we begin moving backward in time, collisions happen relatively quickly at first; but as we continue moving backward, we find that essentially half the expected time is being spent once the lineages have merged down to just two, and these two are searching for a final collision point at the most recent common ancestor. Third, the approximations provide us with a very simple way to build trees according to this process: we simply follow the recipe in Figure 21.21, drawing parallel lines backward in time until the outcome of a coin-flip tells us to pick two of the lines uniformly at random and merge them into one.

Finally, we should note that although we have introduced some approximations into the original formulation of the problem, subsequent work has shown that the final estimates are very close to the exact results one gets through much more intricate analysis [91, 174].

21.9 Exercises

1. Suppose you are studying the spread of a rare disease among the set of people pictured in Figure 21.22. The contacts among these people are as depicted in the network in the figure, with a time interval on each edge showing when the period of contact occurred. We assume that the period of observation runs from time 0 to time 20.

   (a) Suppose that s is the only individual who had the disease at time 0. Which nodes could potentially have acquired the disease by the end of the observation period, at time 20?

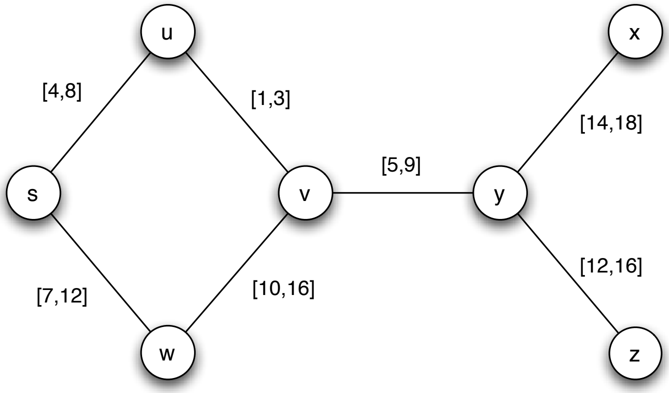


Figure 21.22: Contacts among a set of people, with time intervals showing when the contacts occurred.

(b) Suppose that you find, in fact, that all nodes have the disease at time 20. You're fairly certain that the disease couldn't have been introduced into this group from other sources, and so you suspect instead that a value you're using as the start or end of one of the time intervals is incorrect. Can you find a single number, designating the start or end of one of the time intervals, that you could change so that in the resulting network, it's possible for the disease to have flowed from s to every other node?

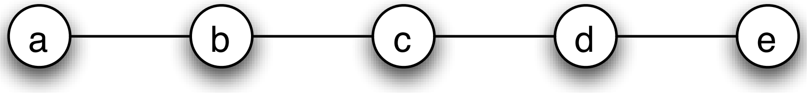


Figure 21.23: A contact graph on five people.

2. Imagine that you know a contact graph on a set of people, but you don't know exactly the times during which contacts happened. Suppose you have a hypothesis that a particular disease passed between certain pairs of people, but not between certain other pairs. (Let's call the first set of pairs *positive*, and the second set of pairs *negative*.) It's natural to ask whether it's possible to find a set of time intervals for the edges that support this hypothesis in a strong sense: they make it possible for the disease to flow between the positive pairs, but not between the negative pairs.

   Let's try this genre of question out on the simple contact graph among five people shown in Figure 21.23.

(a) Can you find time intervals for the edges that make it possible for the disease to flow *from* every node *to* every other node, with the one exception that it is not possible for it to flow from node a to node e? If you think it is possible, describe such a set of time intervals; if you think it is not possible, explain why no such set of time intervals exists.

(b) Can you find time intervals for the edges that make it possible for the disease to flow from a to d and from b to e, but not from a to c? Again, if you think it is possible, describe such a set of time intervals; if you think it is not possible, explain why no such set of time intervals exists.

3. Imagine that you're advising a group of agricultural officials who are investigating measures to control the outbreak of an epidemic in its early stages within a livestock population. On short notice, they are able to try controlling the extent to which the animals come in contact with each other, and they are also able to introduce higher levels of sanitization to reduce the probability that one animal passes the disease to another.

   Both of these measures cost money, and the estimates of the costs are as follows. If the officials spend x dollars controlling the extent to which animals come into contact with each other, then they expect each animal to come into contact with

$$40 - \frac{x}{200,000}$$

   others. If the officials spend y dollars introducing sanitization measures to reduce the probability of transmission, then they expect the probability an infected animal passes it to another animal contact to be

$$.04 - \frac{y}{100,000,000}.$$

   The officials have two million dollars budgeted for this activity. Their current plan is to spend one million on each of the two kinds of measures. Using what you know about epidemics, would you advise them that this is a good use of the available money? If so, why? If not, can you suggest a better way to allocate the money?