

course. Other outbreaks over the past three decades have killed similar numbers in roughly similar circumstances, mostly small, isolated villages with poor health care facilities. The horror stories of these outbreaks are legion: victims coming to the local doctor, complaining of flulike symptoms; crashing and bleeding several days later in the nearest medical clinic; the terrible realization, usually too late, that Ebola has struck; heroic medical workers cut down in the first line of defense; widespread panic; dozens of bleeding bodies discovered in deserted huts; villages left devastated and abandoned; whole regions terrorized. Ebola is truly a monster, a messenger sent straight from Hell.

Ironically, Ebola's tremendous violence is also its one weakness: it is literally too deadly for its own good. Unlike silent, insidious HIV, Ebola exhibits all the subtlety of a train wreck, revealing its true nature in a matter of days and killing shortly thereafter. Furthermore, once the symptoms appear, the victims are so incapacitated and so obviously ill that they have difficulty traveling and can be quarantined with relative ease, thus reducing the virus's ability to spread to new hosts. As a result, the majority of outbreaks have been contained in remote areas near the rain forest and away from major population centers.

Only once, during the second outbreak in 1976, did Ebola make its way to the big city, when a young nurse known only as Mayinga N., infected with the Zaire strain, spent the day wandering around Kinshasa, the capital and largest city of Congo. Fortunately catastrophe was averted by another quirk of the virus: Ebola, at least in its early stages, is not all that contagious. Even when a patient is in a terminal state, hemorrhaging internally and coughing blood-infused mucus into the air, it is generally thought that the virus can only reach new hosts through a break in their skin, or through a permeable membrane like those in the nose or eyes. By the time nurse Mayinga had reached that stage, however, she had already realized her fate and was quarantined in the hospital.

Reading this, you might think that Ebola is just another line item in the seemingly endless litany of horrors afflicting sub-Saharan Africa. And Africa, that most exotic and tragic of continents, is surely

CHAPTER SIX

Epidemics and Failures



THE HOT ZONE

MOST OF US, IT'S PROBABLY FAIR TO SAY, DON'T LOSE A LOT OF sleep over the possibility of catastrophic epidemics. That's probably because most of us have not read *The Hot Zone*, Richard Preston's true story of Ebola, an astonishingly lethal virus that kills its victims in a blood-gushing finale of such heartless fury that only nature could have devised it. Named after the Ebola River, which drains the northern districts of what used to be Zaire but is now the Democratic Republic of Congo, the virus first emerged from its hiding place in the jungles in 1976. It struck first in Sudan and then two months later in Zaire, where it broke out in fifty-five villages almost simultaneously, claiming nearly seven hundred lives that year alone.

Although surprisingly little is known about it, Ebola is thought to have jumped, like HIV, from monkeys to humans and comes in at least three strains of increasing deadliness. A recent outbreak of Ebola in Uganda was of the Sudan strain, which, with a kill rate of a mere 50 percent, is the ninety-pound weakling of the family (Ebola Zaire claims 90 percent of its victims). Even so, 173 people died in the district of Gulu between October 2000 and January 2001 before the outbreak ran its

distant enough that the next plague, if and when it comes, need not affect us any more dramatically than an occasional wince as we flip through the morning newspaper. If *The Hot Zone* has one thing to teach you, however, it is that you can stop relaxing right now. Ebola is a problem not just for Africa but for the whole world. Just as HIV crawled its grisly way down the Kinshasa highway from its birthplace in the jungles and somehow, probably in one of the coastal cities, found Gaetan Dugas—the Canadian flight attendant, better known as *patient zero*—who thereby brought it to the bath houses of San Francisco and introduced AIDS to the Western world, so too could the right chain of events free Ebola from its shackles.

More so even than Preston's vivid descriptions of Ebola-induced death, it is the potential for a global explosion of the virus that is the most disturbing aspect of his account. Over the last century, not only have we humans intruded deeply into the ancient ecologies of the African rain forests, where the most deadly viruses lie in wait, but also we have built an international system of transportation networks that can transmit an infectious disease to the world's metropolises and power centers within a few days—less time, it so happens, than Ebola's incubation period. Preston even says of one of his doomed characters, as he sits on a small plane to Nairobi vomiting bagloads of black blood, "Charles Monet and the life form inside him had entered the net."

The prospect of Ebola showing up in the local shopping mall is almost too horrible to contemplate, but after reading *The Hot Zone*, you're almost amazed that it hasn't. In fact, the main subplot of the book describes the outbreak of a third strain of Ebola among a population of monkeys at an Army research lab in Reston, Virginia, just outside Washington, D.C. The virus, now identified as Ebola Reston, turned out to be harmless to humans but spectacularly lethal to the poor monkeys, all of whom died. But Ebola Reston is so similar to Ebola Zaire that none of the standard tests at the time could tell them apart, and for a few nail-biting days, Zaire is what the scientists and animal handlers who had been exposed to it thought it was. Had it actually been

the Zaire strain—and it is pure dumb luck that it wasn't—then we would all know a lot more about Ebola today than we do.

VIRUSES IN THE INTERNET

THESE DAYS BIOLOGICAL VIRUSES ARE NOT THE ONLY POTENTIAL source of epidemics, as Claire Swire discovered much to her chagrin right before Christmas 2000. Claire Swire is a young Englishwoman who a few days earlier had apparently had a brief affair with a young Englishman named Bradley Chait. Being a modern woman, she sent him an e-mail the following day, complimenting him in a way that Chait found so appealing, he decided to share it with his friends. Only his best friends, mind you—only six of them. But they apparently found the compliment so entertaining that each of them forwarded it to several of *their* nearest and dearest, many of whom, it turns out, felt the same way. And so it went, this little e-mail, with Chait's one-line amendment, "That's a nice compliment from a lass," around and around the world, amusing roughly 7 million readers in a matter of days. *Seven million!* Poor Claire had to go into hiding to avoid the ensuing press frenzy, and Chait was "disciplined" by his law firm for unauthorized use of his e-mail account (as if people don't send personal e-mail from work all the time). It's a silly story maybe, but a fine example of the power of exponential growth, especially when mixed with the near costless transfer of information afforded by the Internet. And on this topic, there are plenty of serious things to say.

Viruses, both human and computer, essentially perform a version of what we have been calling a *broadcast search* throughout a network. Broadcast searches, as discussed in chapter 5, represent the most efficient way of starting from any given node and finding every other one by systematically branching out from each newly connected node to each of its unexplored neighbors. When a disease embarks on a "search," however, it isn't looking for anything in particular—it is

simply seeking to spread itself as far and wide as possible. So “efficiency” for an infectious entity like a virus generally carries with it connotations of mayhem. The more contagious a virus is, and the longer it can keep the host in an infectious state, the more efficient it is at searching. Ebola, therefore, is more efficient than HIV in that it is significantly more infectious (HIV-infected patients don’t vomit blood in the emergency room), but is less efficient in that it kills so quickly. And both HIV and Ebola are far less efficient than the influenza virus, which not only keeps its hosts alive for much longer but also is able to spread via airborne particles. To put the importance of disease efficiency in perspective, if Ebola were contagious via airborne transport, modern civilization might well have come to an end sometime in the late 1970s.

As much as we should be concerned about the possibility of a human “slate wiper,” as Preston dubs truly devastating plagues, in terms of efficiency alone computer viruses are far more troublesome than human ones. A virus—whether human or computer—can be regarded as little more than a set of instructions for reproducing itself, using material from the host as its building blocks. In humans the immune system screens out foreign and possibly dangerous sets of instructions, but computers don’t generally possess immune systems. In essence, the function of a computer is to execute instructions as efficiently as possible, regardless of where the instructions came from. So they are considerably more vulnerable to malicious bits of code than are people. And although a worldwide computer epidemic might not signal the end of civilization, it could still exert a significant economic toll. No such event has occurred yet, but we have already experienced some disquieting tremors. In the last few years of the twentieth century, even before Y2K turned out to be the biggest anticlimax of the millennium, a series of computer virus outbreaks caused a significant level of disruption and inconvenience to hundreds of thousands of users around the world. Government agencies, large corporations, and even the usually ambivalent public began to sit up and take notice.

Computer viruses have been with us for decades, so why have we only recently started to experience them on a global scale? The answer, as it is for so many questions about the second half of the 1990s, is the Internet. Before the Internet, viruses circulated and computer users experienced occasional difficulties. But back then, pretty much the only way to transmit a virus from one machine to another was via a floppy disk, which had to be physically inserted into the machine. Certainly it was possible for the contaminated disk to circulate among many computers, and once a computer was infected, saving related files onto an uninfected floppy would infect that disk as well. So the potential for exponential growth clearly existed, but the largely manual nature of spread—like Ebola’s requirement for a break in the skin—generally reduced the efficiency of the virus enough that small outbreaks did not become full-fledged epidemics.

The Internet, particularly e-mail, has changed all that, as the world began to understand in March 1999 with the arrival of the Melissa virus. Although Melissa was generally referred to as a virus (or a bug), it actually had a lot in common with another type of malicious code known as a worm. Worms wreak havoc not so much on individual computers as on networks of computers. They replicate and transmit themselves in large numbers from machine to machine without being activated by a user. Melissa, which at the time was the fastest-spreading virus that anyone had ever seen, arrived in the form of an e-mail with the subject line, “Important message from <name>,” where <name> was that of the user sending the message. The body of the message said, “Here is the document you asked for . . . don’t show anyone else,;-)” and a Microsoft Word document called list.doc was attached. If the attachment was opened, Melissa’s macro would automatically mail copies of itself to the first fifty addresses in the user’s e-mail address book. If any of the addresses happened to be a mailing list, everyone on the list would get the virus.

The results were quite dramatic. First detected on Friday, March 26, Melissa had spread all over the globe within hours, and by Monday morning it had infected over one hundred thousand computers in three

hundred organizations, bombarding some sites with so many messages (in one case, thirty-two thousand messages in forty-five minutes!) that they were forced to take their mail systems off-line. It could have been a lot worse, however. Melissa not only was relatively benign—its worst action was to insert a harmless reference to *The Simpsons* into an open document if the minute of the hour matched the day of the month—but also was only able to propagate via the Microsoft Outlook mail program. Users without Outlook could still receive the virus but were unable to pass it on, a distinction that has important consequences for the likelihood of a truly devastating global virus (and possibly even for the Microsoft corporation itself) as we will discuss later. First, however, we have to learn a thing or two about the mathematics of infectious disease. In particular, we need to understand better the conditions under which a small outbreak of a disease becomes an epidemic.

THE MATHEMATICS OF EPIDEMICS

MODERN MATHEMATICAL EPIDEMIOLOGY WAS BORN OVER SEVENTY years ago with the introduction of the *SIR model*, which was formulated by two mathematicians, William Kermack and A. G. McKendrick, and is still the basic framework around which most infectious disease models are constructed. The letters in the acronym represent the three primary states (illustrated in Figure 6.1) that any member of a population can occupy with respect to a disease: *susceptible*, meaning that the individual is vulnerable to infection but has not yet been infected; *infectious*, implying that the individual not only is infected but also can infect others; and *removed*, implying that the individual either has recovered or has otherwise ceased to pose any further threat (possibly by dying). New infections can only occur when an infected individual, often called an *infective*, comes into direct contact with a *susceptible*. At that point, the susceptible can become infected, with a probability that depends on the infectiousness of the disease and the characteristics of the susceptible—(some people, clearly, are more susceptible than others).

Obviously, who comes into contact with whom will depend on the network of associations in the population. To complete the model, therefore, we must make some assumptions about that network. The standard version of the model, for example, assumes that interactions between members of the three subpopulations occur

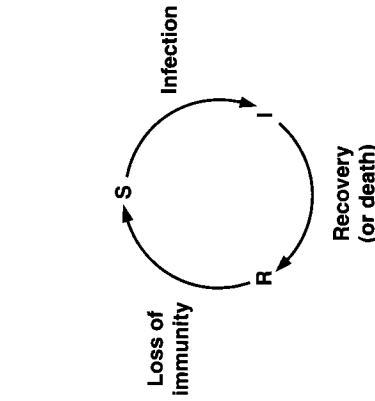


Figure 6.1. The three states of the SIR model. Each member of the population can be susceptible, infected, or removed. Susceptible individuals can become infected by interacting with infectives. Infectives can either recover or die, thus ceasing to take part in the dynamics. If they recover, they might become susceptible again through loss of immunity.

purely at random, as if all the members of the population were being stirred in a large vat, like the one in Figure 6.2. As the image of the vat suggests, pure randomness isn't a very good proxy for human interactions, but it certainly simplifies the analysis considerably. In the SIR model, the randomness assumption implies that the probability of an infective meeting a susceptible is determined solely by the size of the infected and susceptible populations—in a vat, there is no population structure to speak of. The problem still isn't trivial, but now at least it is possible to write down a set of equations whose solutions depend only on the size of the initial outbreak and a few parameters of the disease itself, like its infectiousness and the recovery rate.

According to the model, when an epidemic does occur, it should follow a predictable course known to mathematicians as *logistic growth*. As Figure 6.3 indicates schematically, each infection requires the participation of both an infected and a susceptible individual. Hence, the

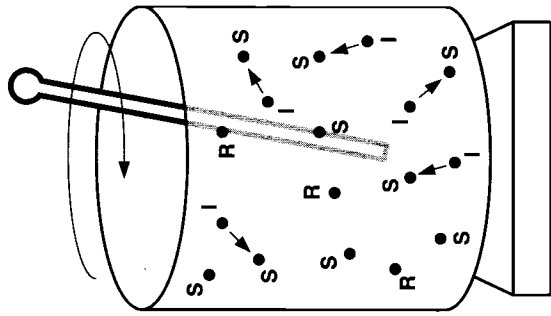


Figure 6.2. In the classical version of the SIR model, interactions are assumed to be purely random. One way to think of random interactions is as individuals being mixed together in a large vat. The main consequence of the random mixing assumption is that interaction probabilities depend only on the relative population sizes, a feature that greatly simplifies analysis.

rate at which new infections can be generated depends on the size of both populations. When the disease is in its early stages, the infected population is small, and therefore so is the rate of new infections—as the top diagram in Figure 6.3 shows, there just aren't enough infectives to cause much damage. This *slow-growth phase* is also the most effective stage in which to prevent an epidemic, as even a few averted infections can drive the disease back into remission. Unfortunately, an epidemic in its early stage can be hard to distinguish from a random grouping of unrelated cases, especially if public health authorities are poorly coordinated or reluctant to admit that they have a problem.

By the time the density of infectives becomes too great to be overlooked or ignored, the epidemic has typically entered the *explosive phase* of logistic growth (middle diagram in Figure 6.3). Now there are

many infected and many susceptible individuals, so the rate at which new infections occur is maximized. Epidemics in the midst of explosive growth are essentially impossible to stop, as British farmers witnessed in 2001 when foot-and-mouth disease raged for half a year throughout most of England and parts of Scotland. When the epidemic was detected in mid-February, only three weeks after the first cases had occurred, forty-three farms had already been affected. That might seem like a lot of farms, but the epidemic was still in the initial,

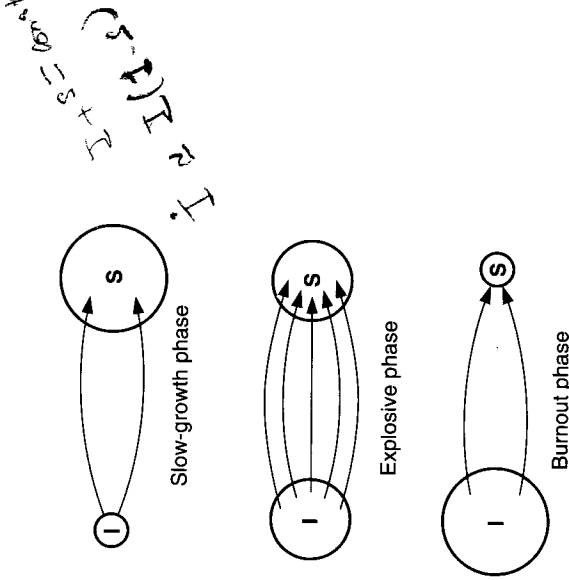


Figure 6.3. In logistic growth, the rate of new infections depends on the size of the susceptible and infected populations. When either population is small (top and bottom diagrams), new infections are rare. But when both populations are intermediate in size (middle diagram), infection rates are maximized.

slow-growth phase. By September, the number of farms suspected of infection had grown to over nine thousand, despite the preventative slaughter of nearly 4 million sheep and cattle.

Eventually, however, even the most out-of-control epidemics come to an end, if for no other reason than they burn themselves out. Because there are only so many people (or in the case of foot-and-mouth dis-

ease, animals) who can be infected, susceptible targets become harder and harder to find, and the trajectory of the disease flattens off again. This is the *burnout phase* of logistic growth. In the foot-and-mouth epidemic, this self-limiting process was accentuated by the effective quarantine of farmlands and the massive extermination of animals (only about two thousand actual cases of the disease were ever detected, a tiny percentage of the number killed). From start to finish, therefore, the course of an epidemic displays a characteristically S-shaped curve—like the one in Figure 6.4. That the main features of this trajectory—slow growth, explosion, and burnout—are explicable in terms of the logistic growth model suggests that the forces governing an epidemic, when it occurs, are fundamentally quite simple.

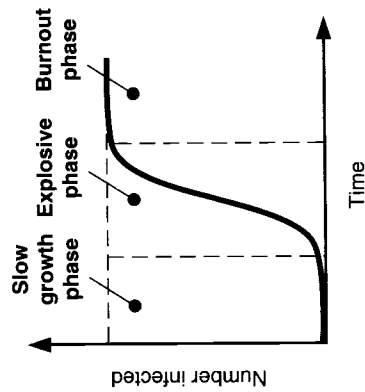


Figure 6.4. Logistic growth, displaying the slow-growth phase, explosive phase, and burnout phase.

But epidemics do not always occur. In fact, most outbreaks of disease either are contained by human intervention or (much more often) burn themselves out before infecting more than a tiny fraction of the population. As terrifying as it was, the Ebola outbreak of 2000, for example, doesn't qualify as a genuine epidemic. Although 173 victims is a significant number in absolute terms, the outbreak remained confined to a geographically localized collection of villages, never seriously threatening the bulk of the potentially vulnerable population. The foot-and-mouth epidemic of 2001, by contrast, affected almost the entire country. In terms of the SIR model, stopping an epidemic is roughly equivalent to preventing it from reaching the explosive growth

phase of Figure 6.4, which in turn implies focusing not on the size of the initial outbreak but on its *rate of growth*. The key measurement of a disease in this respect is its *reproduction rate*, the average number of new infectives generated by each currently infected individual.

The mathematical condition for an epidemic is that the reproduction rate of the disease has to be greater than one. If the reproduction rate is kept below one, then infectives are removed from the population faster than new ones are generated, and the disease will die out without becoming an epidemic. But if the reproduction rate exceeds one, then the disease increases not only its spread but also the speed with which it will continue to spread, and explosive growth inevitably commences. The *knife edge* between these two conditions, where a single host passes on its burden to precisely one single new host, is called the *threshold* of an epidemic. Preventing an epidemic amounts to keeping the reproduction rate below its threshold.

In the classic SIR model, where population structure is ignored, the reproduction rate, and therefore the epidemic threshold, is determined entirely by the properties of the disease itself (its infectiousness, and the speed with which infectives recover or die) and by the size of the susceptible population with which its hosts can interact. Thus, safe-sex practices have constrained the HIV epidemic in some parts of the world by targeting its infection rate, while the widespread extermination of animals in Britain during the foot-and-mouth epidemic very likely reduced its severity by limiting the effective size of its susceptible population.

That the threshold reproduction rate in the classical model should be exactly one turns out to be one of those deep convergences that makes mathematics so interesting. The epidemic threshold is, in fact, exactly analogous to the critical point at which a giant component appears in a random network (see chapter 2), where the reproduction rate is mathematically identical to the average number of network neighbors. And the size of the infected population as a function of the reproduction rate (Figure 6.5) is exactly analogous with the size of the giant component in Figure 2.2. The onset of an epidemic, in other words, occurs when the disease passes through exactly the same phase

transition that Erdős and Rényi discovered in their apparently unrelated problem about communication networks. This remarkable similarity, however, also suggests an obvious criticism. If we rejected random graph models as realistic representations of real-world networks, social or otherwise, should we not also reject any conclusions

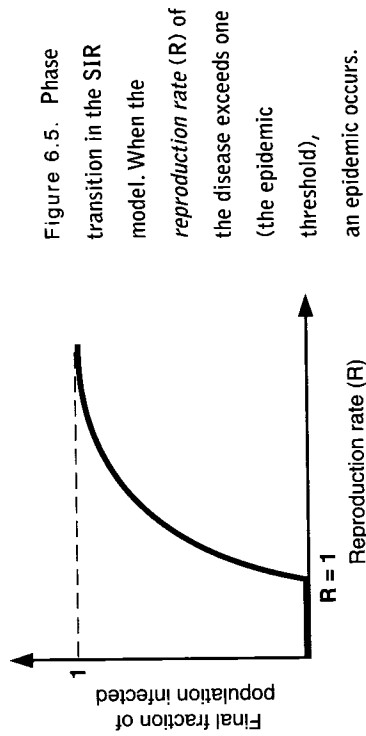


Figure 6.5. Phase transition in the SIR model. When the reproduction rate (R) of the disease exceeds one (the epidemic threshold), an epidemic occurs.

about epidemics that are based on the same assumptions? The dependence of the reproduction rate on the size of the susceptible population alone, for example, doesn't account for any of the features of social or network structure that might be useful in combating an epidemic. As we will see, some lessons of the classical model hold up even in the complex world of networks, but new, network-oriented lessons have to be learned as well.

E P I D E M I C S I N A S M A L L W O R L D

STEVE AND I, REMEMBER, WERE INTERESTED IN DYNAMICS FROM the start. After all, we had gotten ourselves into the network business originally because we were interested in the dynamics of coupled oscillators—the crickets. So once we had some network models to play with, we naturally wondered how different dynamical systems might behave on them. The first such system that we tried to understand was the Kuramoto oscillator model, from chapter 1, on which Steve had done so

much work earlier in his career. Unfortunately, as simple as the Kuramoto model is, its behavior on a small-world network was still too complicated for us to make sense of it (a statement that remains true even several years later). So we started looking for a simpler kind of dynamics, and once again, Steve's biological interests came in handy. "The SIR model is about the simplest kind of nonlinear dynamics I can think of," he said one day in his office, "and I'm pretty sure that no one has really thought about the SIR model located on a network—at least not a network anything like this. Why don't we try that?"

So we did, but this time I did some homework first. Sure enough, while the basic SIR model had been generalized in many ways to include the idiosyncrasies of particular diseases and the varying susceptibilities of different demographic groups, nothing like small-world networks had come up in the literature. That much was encouraging, as was the deep equivalence between the classic SIR model and the connectivity of a random graph. Whatever the behavior of a disease in a general small-world network, we could be sure that it had to resemble the classical SIR behavior in the limit where all links had been randomly rewired (as in the right panel of Figure 3.6). So not only did we have a network model, which by this stage we understood reasonably well, we also had a well-established benchmark against which to compare our results.

The first natural comparison to make against the random limit was for a disease spreading on a one-dimensional lattice—the ordered end of the small-world spectrum, and the left panel of Figure 3.6. In a lattice, as discussed in chapter 3, the links between nodes are highly clustered, implying that a spreading disease is continually being forced by the network back into the already infected population. As displayed in Figure 6.6, in a one-dimensional lattice, a growing cluster of infectives actually consists of two kinds of nodes—those in the interior of the cluster (who cannot infect any susceptibles) and those on the boundary, or *disease front*. No matter how large the infected population is, the size of the disease front remains fixed; hence, the *per capita* rate of growth of the infected population inevitably decreases as the infection spreads.

Thus, a lattice presents a very different context for an epidemic than does the random mixing assumption above. It also makes the reproduction rate difficult to compute, so we decided to compare the results for our different networks directly in terms of infectiousness. And the difference was striking. As shown in Figure 6.7, the same disease,

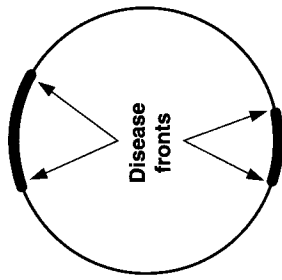


Figure 6.6. On a ring lattice, the disease front (where infectives and susceptibles interact) is fixed. As the size of the infected population increases, more infectives lie in its interior, where they cannot reach new susceptibles. Hence, diseases spread slowly on lattices.

spreading in a lattice, tends to infect far fewer people than it does in a random graph, and there is no longer any clear threshold. The take-home message is that when diseases are confined to spread in only a limited number of dimensions—even, say, by the two-dimensional geography of the land—only the most infectious diseases will develop into true epidemics. And even then, they will be slow, creeping epidemics, rather than explosions, giving public health authorities time to respond and a well-defined area on which to focus.

An example of just such a creeping epidemic is the black death that swept across Europe in the fourteenth century, wiping out about one-fourth of the entire population. As mind boggling a statistic as that is, an epidemic like the plague probably couldn't happen today—at least not in the industrialized world. As the map in Figure 6.8 shows, the plague started in a single town in southern Italy (where it is thought to have arrived on an infected ship from China) and then spread like a ripple along the surface of a pond after a stone is dropped. Because the disease was transported mostly by rats infested with plague-carrying

fleas it took three years, from 1347 to 1350, for the front to propagate across Europe. Neither medical science nor public health services at that time were able to prevent the plague's relentless progress, so its relatively slow speed didn't make much difference ultimately. But in the modern world, any disease forced to travel by such slow and inefficient means could be identified and contained.

Unfortunately, diseases today have much better mechanisms for transport than scurrying rats. And no sooner did we permit even a small fraction of random links into our network models, than the relative stability of the lattice model splintered apart. To see this effect,

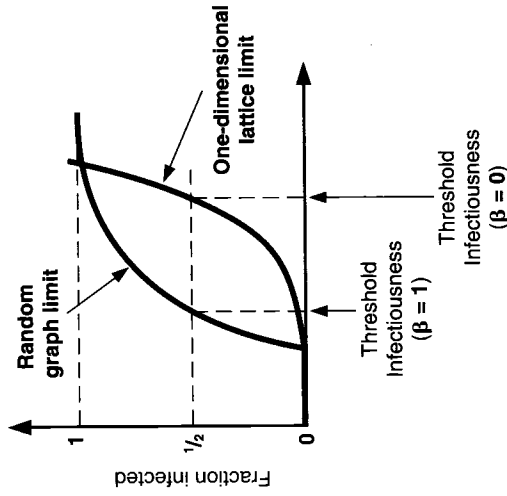


Figure 6.7. The fraction infected versus infectiousness for the random graph limit ($\beta = 1$) and lattice limit ($\beta = 0$) of the beta model from chapter 3. The value for the threshold infectiousness represents the infectiousness required for one-half of the population to become infected.

consider the horizontal line drawn halfway up across Figure 6.7. The points at which the two infection curves intersect the line represent the values of the disease infectiousness at which that fraction of the population is infected (in the figure, the fraction is one-half, but we could have chosen some other value). Call this value the *threshold infectiousness* (remember we can no longer use the reproduction rate to define the threshold for an epidemic, so we use a fixed fraction of the population instead), and then ask how it varies with the fraction of random shortcuts in the network. As we see in Figure 6.9, the threshold infectious-

ness starts high—the disease must be highly infectious in order to contaminate a large population—but then drops rapidly. More importantly, it approaches the worst-case scenario of a completely random network while the network itself is still far from random.

This observation might help explain why epidemics such as the

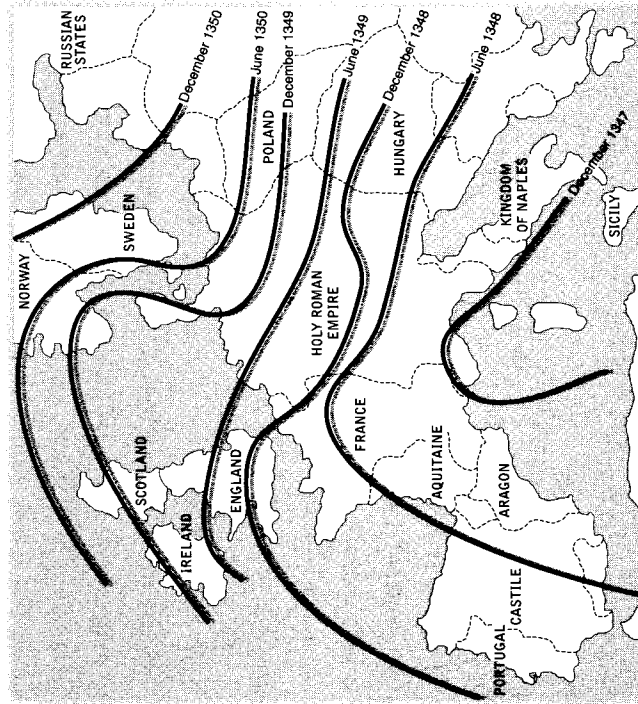


Figure 6.8. Map of the progression of the plague (black death) in Europe between 1347 and 1350.

foot-and-mouth disease epidemic in Britain can explode so rapidly. Because foot-and-mouth disease spreads between animals either through direct contact or indirectly via wind-blown droplets excreted from symptomatic animals and in virus-laden soil, one might expect any initial outbreak to spread only along the two-dimensional geography of the English countryside, much like the plague did seven hundred years earlier. However, the combination of modern transportation, modern livestock markets (in which animals from geographically dispersed forms are exchanged, or simply come into physical contact), and recreational

hikers carrying infected soil on their boots has broken the constraints of geography. As a consequence, British sheep and cattle farms are linked by a network of transportation systems that can move infected animals (and people) anywhere in the nation overnight. And because these links are, for all intents and purposes, random, the virus only needed to find a few of them in order to launch itself into fresh territory. An important early problem in combating the epidemic, for example, was that the forty-three farms on which foot-and-mouth disease was first detected were not neighboring farms. Hence, the virus had to be fought on many fronts simultaneously, with more being added every day.

That the results of the random mixing model turn out to be so easily replicated even in highly clustered networks is not good news for the

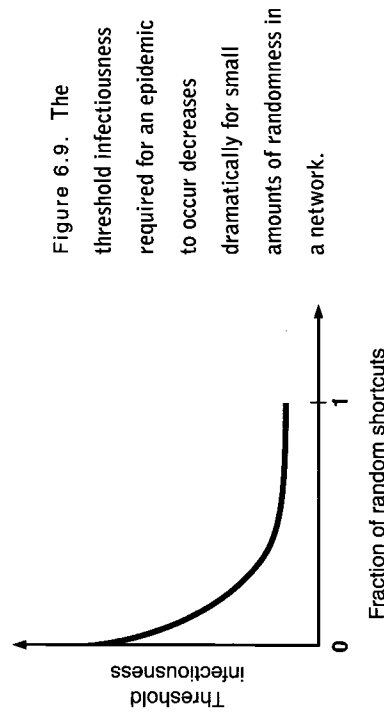


Figure 6.9. The threshold infectiousness required for an epidemic to occur decreases dramatically for small amounts of randomness in a network.

world. If diseases really do spread on small-world networks, then it would seem that we are continually facing a worst-case scenario. Even more troubling, because very few people ever have more than local information about their networks, it can be very hard for public health authorities to make individuals understand the immediacy of an apparently distant threat and thus change their behavior. AIDS is a good example of this problem. For more than a decade after the AIDS epidemic was first identified, HIV infection was generally considered to be confined to only a few, quite specific communities—gay men, prostitutes, and intravenous drug users. So if person X didn't have sex with anyone in these three categories, and none of his or her sexual partners

did either, then person X was safe, right? Wrong! What is obvious to us now that we have seen the virus infect almost entire nations in southern Africa is that in the small world of sexual networks, even an apparently distant danger must be taken seriously. Particularly troubling is the thought that HIV was able to breach its initial boundaries at least in part because of the perception that it couldn't.

The phrase "think globally, act locally" is therefore nowhere more appropriate than to the prevention of epidemics. Remember that infectious diseases, unlike the search problems of the previous chapter, conduct what we have called broadcast searches. So if there is a short path through a network of contacts between an infective and a susceptible, it doesn't matter if either person knows it's there or even whether they could find it if they wanted to. Unless the disease is stopped somehow, it will find the path because it is blindly probing the network for every path. And unlike Gnutella users or Mrs. Forrest's sixth-grade class from the previous chapter, it is quite happy to overload the entire network with copies of itself—that's what infectious diseases do. That our perception of risk from an infectious disease, whether it be HIV, Ebola, or even, say, the West Nile virus, should be so out of kilter with the actuality of disease transmission is definitely a cause for concern.

The situation is not all gloom and doom, however. As stated earlier, outbreaks of disease typically do not become epidemics, and in this respect small-world networks have something encouraging to teach us. On a small-world network, the key to explosive growth of a disease is the shortcuts. Diseases don't spread very effectively on lattices, and although small-world networks exhibit some important features of random graphs, they still share with lattices the property that locally, most contacts are highly clustered. So *locally*, the growth of a disease behaves very much like it does on a lattice: infected individuals interact mostly with other already infected individuals, preventing the disease from spreading rapidly into the susceptible population. Only when the disease cluster reaches a shortcut—whether that be an Ebola victim getting on a plane, or a truckload of cattle infected with foot-and-mouth driving up the M1—does it start to display the worst-case, random mixing

behavior. So unlike epidemics on a random graph, epidemics in a small-world network have to survive first through a slow-growth phase, during which they are most vulnerable. And the lower the density of shortcuts, the longer this slow-growth phase will last.

A network-oriented strategy for preventing epidemics, therefore, would not only attempt to reduce infection rates in an overall sense, it would focus particularly on likely sources of shortcuts. Interestingly enough, the needle exchange program that has been effective in reducing the spread of HIV among intravenous drug users exhibits both these features. Removing dirty needles from circulation eliminates one mechanism by which HIV can spread, thus reducing the overall rate of infection. But it also works by virtue of the *particular* infections that it prevents. Dirty needles are shared not only among friends but also by complete strangers, who might pick up and reuse a discarded needle. In other words, reused needles are a source of random connections in the disease network. Just as the ban on animal movements and the closure of country footpaths throughout England in 2001 reduced the potential for long-range shortcuts, eliminating needles from the system closes off one avenue of escape from the slow-growth phase of an epidemic, giving health authorities a better chance of catching up with the disease.

Thinking about the structure of networks may also explain other subtleties in the spread of a disease that would not be apparent in the absence of a network-oriented approach. Recently the Spanish physicist Romualdo Pastor-Satorras and the Italian physicist Alessandro Vespignani pointed out one such feature of real-world computer viruses that classical SIR models have trouble explaining. After studying prevalence data available at a popular on-line virus bulletin, they concluded that most viruses exhibit a peculiar combination of long-term and low-level persistence "in the wild." The combination is peculiar because according to the standard SIR model, every virus must either generate an epidemic (in which case some significant fraction of the population will be infected) or quickly burn itself out. In other words, either it explodes or it doesn't. But unless it happens to have a repro-

duction rate of exactly one, the critical point of the phase transition in Figure 6.5, it can't just drift around failing to do either. By contrast, many of the 814 viruses whose timelines were recorded on the virus bulletin appeared to do precisely that. Some of them had been floating around for years, despite the availability of antivirus software usually within days or weeks after their initial detection.

Pastor-Satorras and Vespignani proposed an explanation that explicitly included features of the e-mail network through which they hypothesized the viruses were spreading. Taking Barabási and Albert's scale-free model as a proxy for the structure of e-mail networks—an assumption that was supported (albeit inconclusively) a year later in a report by a German team of physicists—the two physicists showed that when spreading on scale-free networks, viruses don't display the same threshold behavior that they do in the standard model. Instead, as in Figure 6.10, the fraction of the population infected tends to grow continuously from zero as the infectiousness of the disease increases. In a

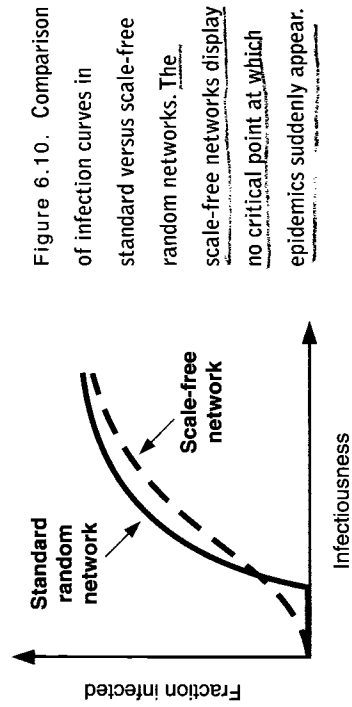


Figure 6.10. Comparison of infection curves in standard versus scale-free random networks. The scale-free networks display no critical point at which epidemics suddenly appear.

scale-free e-mail network, most nodes have only a few links, which is to say that most people send e-mail to only a few others on a regular basis. But a small fraction of e-mail users have very extensive address books, containing a thousand names or more, and are apparently diligent enough to keep up with them all! It is this minority that Pastor-Satorras and Vespignani hypothesized is more or less responsible for the

long-term persistence of viruses—only one of them needs to become infected with a virus once in a while for it to continue to circulate at measurable levels throughout the population as a whole.

Apparently then, even the simplest features of real-world networks, like local clustering and scale-free degree distributions, can have important consequences for the spread of disease and, most important, the conditions governing epidemics. The study of disease models is, therefore, an important subfield in the new science of networks. In a world where several tens of millions of people are now infected with HIV and where the prevalence varies, even within Africa, from less than 2 percent to more than one third of a country's population, the importance of understanding the spread of infectious disease in networks cannot be overstated. Much work remains to be done, but already some promising directions are appearing in the network literature. And while the SIR model remains central to this effort, the physicists, true to form, have started tackling the problem in their own way. In particular, they have introduced to the study of epidemics, a set of techniques that goes under the general label of *percolation theory*.

PERCOLATION MODELS OF DISEASE

HISTORICALLY, PERCOLATION THEORY DATES BACK TO WORLD War II, when Paul Flory and his collaborator Walter Stockmayer used it to describe the *gelation* of polymers. If you have ever boiled an egg, then you are familiar with some aspects of polymer gelation. As the egg is heated, the polymers in the egg white link up and bond to each other, one pair at a time. Then at some critical point, the white undergoes a sudden, apparently spontaneous transition, called *gelation*, in which a very large number of branching polymers suddenly become bound together in a single, coherent cluster that spans the entire egg. In break-fast terms, before gelation, the egg is liquid; after gelation, it is solid. The first success of percolation theory was Flory and Stockmayer's

explanation of how this transition could happen almost instantaneously, not slowly and incrementally as one might have expected. Although it was originally developed to answer questions in organic chemistry, percolation theory has subsequently proved a useful way of thinking about all sorts of problems, from the sizes of forest fires, to yields from underground oilfields, to the electrical conductivity of composite materials. More recently, it has also been used to think about the spread of disease.

In late 1998, not long after I had arrived at the Santa Fe Institute, I had started talking with Mark about the disease-spreading work that I had done with Steve the year before. Based on a simple SIR model, Steve and I had been able to make some conclusions about the dependence of the epidemic threshold on the density of random shortcuts. But we hadn't been able to understand exactly how the mechanism worked, or how the effect of random shortcuts varied with the density of the network. Since then I had been teaching myself some of the basics of percolation theory, which seemed a natural way to go about asking many of the same questions. And Mark, being an expert in statistical physics, was the obvious person to ask. As I soon learned was typical with Mark, once he got interested in the problem, it didn't take very long for results to follow.

Imagine a very large population of individuals (*sites* in percolation terminology) connected to one another by a network of ties (*bonds*) along which a disease might be transmitted. Each site in the network is susceptible or not, with some probability, called the *occupation probability*, and each bond is either *open* or *closed* with a probability that is equivalent to the infectiousness of the disease. The result looks something like the diagrams in Figure 6.11 (although for much larger networks), where the disease can be thought of as an imaginary fluid pumped out of a source site. Starting at the source, the disease will always "flow" along any open bonds that it encounters, spreading from one susceptible site to another until no more open bonds can be accessed to new susceptible sites. The group of sites that can be reached in this fashion from a randomly selected start point is called a *cluster*, where the entry of a disease into a given cluster necessarily implies that all the sites in that cluster become infected also.

In the left diagram of Figure 6.11, the occupation probability is high and many bonds are open, implying a highly infectious disease to which most of the population is susceptible. In this condition, the largest cluster spans almost the entire network, thus implying that an outbreak at a random location in the network; would be expected to

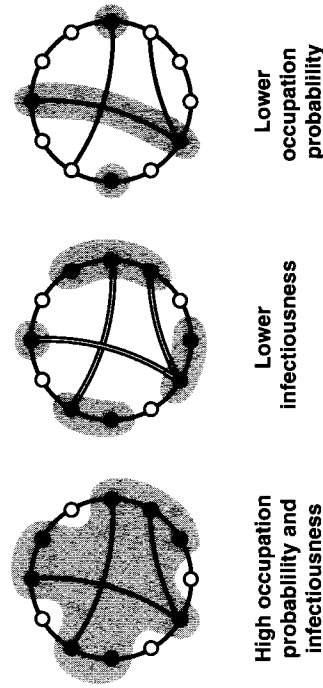


Figure 6.11. Percolation on a network. Solid circles (links) correspond to occupied (open) sites (bonds). Connected clusters are shaded.

spread widely. In the other two diagrams, by contrast, either the infectiousness (middle diagram) or the occupation probability (right diagram) is low, implying that disease outbreaks will be small and localized, no matter where they occur. In between these extremes lies a complicated continuum of possibilities in which clusters of all sizes can exist simultaneously, and the extent to which a disease spreads is determined by the size of the particular cluster in which it originates. The main objectives of percolation theory are to characterize this distribution of cluster sizes and to determine how it depends on the various parameters in the problem.

In the language of physicists, the possibility of an epidemic depends on the existence of what is called a *percolating cluster*—a single cluster of susceptible sites (connected by open bonds) that permeates the entire population. In the absence of a percolating cluster, we would still see outbreaks, but they would be small and localized. However, a dis-

ease that starts somewhere on a percolating cluster, instead of dying out, will spread throughout even a very large network. The point at which a percolation cluster appears, usually referred to as *percolation*, turns out to be exactly analogous to Flory and Stockmayer's gelation in polymers. It is also equivalent to the epidemic threshold in SIR models at which the reproduction rate of the disease first exceeds one (and therefore, by association, the connectivity transition of a random graph). As Figure 6.12 shows, below the threshold, the size of the

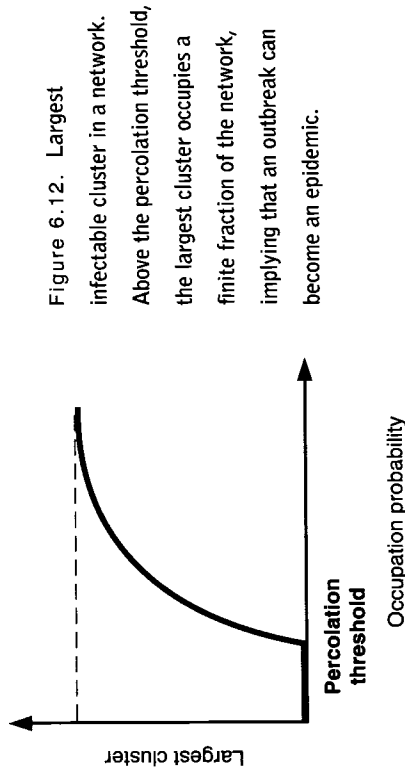


Figure 6.12. Largest infected cluster in a network. Above the percolation threshold, the largest cluster occupies a finite fraction of the network, implying that an outbreak can become an epidemic.

largest cluster, when viewed as a fraction of the whole population, is negligible. But as the critical point is reached, we observe the sudden and dramatic appearance of a percolating cluster—apparently out of nowhere—through which the disease can spread uninhibited.

The distance through a network that a disease will typically spread before burning itself out is equivalent to what physicists call the *correlation length*, a term we encountered in chapter 2 in the context of global coordination. There, the divergence of the correlation length implied the system had entered a critical state in which even local perturbations could propagate globally. Much the same result is true in percolation models of disease spreading. Right at the percolation transition, the correlation length becomes effectively infinite, implying that even very distant nodes can infect one another. What Mark and I fig-

ured out, in the case of small-world networks, was how the correlation length depended on the fraction of random shortcuts. In agreement with the crude results that Steve and I had gotten almost two years earlier, Mark and I showed that even a small fraction of random shortcuts could alter the correlation length dramatically. But now, by solving for the conditions under which the correlation length diverged, we could determine the position of the percolation transition—and thus the epidemic threshold—precisely.

NETWORKS, VIRUSES, AND MICROSOFT

THIS RESULT WAS A PROMISING START AND DEMONSTRATED THAT for some problems at least, epidemics can be better understood by using a percolation approach than with the standard SIR model. Unfortunately, percolation on realistic networks is a difficult (and unsolved) problem, and further progress proved hard to come by. To keep the analysis manageable, for example, most percolation models either assume that all sites in the network are susceptible, and focus on the bonds (this is called *bond percolation*), or assume that all bonds are open, and focus on the sites (*site percolation*). Roughly the same methods work for both kinds of percolation, and in many respects, they behave in a similar fashion. Mark and I, for instance, studied the site percolation version, but shortly thereafter, Mark and another Santa Fe physicist, Cris Moore, extended the results to bond percolation. In some respects, however, site and bond percolation differ significantly, occasionally yielding quite different predictions for the likelihood of an epidemic.

Before racing ahead with the analysis, therefore, one has to think carefully about which version—bond percolation or site percolation—best captures the nature of the disease in question. In the case of a virus like Ebola, for example, it seems reasonable to assume that all people are susceptible, and to focus on the extent to which they can infect each other. Therefore, the relevant formulation of an Ebola-related percola-

tion problem would be bond percolation. Computer viruses like the Melissa bug, however, will generally pass between any susceptible computer and any other computer (all bonds are effectively open), but not all computers are susceptible. So a percolation model of a computer virus probably ought to be of the site percolation variety. Taking the Melissa bug as an example, only a certain fraction of computers in the world are susceptible to the virus because it can only spread via the Microsoft Outlook e-mail program, and not everybody uses Outlook.

Unfortunately for Microsoft users, so many computers run Outlook that the largest connected cluster of them almost certainly percolates. If it didn't, in fact, we wouldn't see global viral outbreaks like Melissa and its protégés, the Love Letter and Anna Kournikova viruses. Universal software compatibility clearly confers some significant benefits on individual users. But from the perspective of system vulnerability, when everybody has the same software, everybody also has the same weaknesses. And every piece of software has weaknesses, especially large, complex operating systems like Microsoft's. In a way, the only amazing thing about Melissa-style outbreaks is that they haven't happened more often. And if they do start to happen more often—if Microsoft software acquires the reputation for persistent vulnerability—then large corporations, and even individuals who cannot afford to have their computers put out of action every time a new virus appears anywhere in the world, may start to look for alternatives.

What should Microsoft do? The obvious approach is to make its products as resilient as possible to attack by any wormlike virus, and in the event of an outbreak, make antivirus software available as quickly as possible. These measures have the effect of reducing the occupation probability of the network, thus shrinking and possibly even eliminating the percolating cluster altogether. But if massive corporations like Microsoft, who are the natural targets of any hacker desiring fame and glory, want to protect their customers and their market share, they may also have to think a little more radically. One solution might be to switch from a single integrated product line to several different products that are developed separately and that are designed not to be entirely compatible.

From a conventional software point of view, one that emphasizes compatibility and economies of scale, deliberately disintegrating a product line might seem like a crazy idea. But in the long term (and the long term may not be that long), a proliferation of nonidentical products would reduce the number of computers susceptible to any particular virus, rendering the system as a whole dramatically less vulnerable to the largest of viral outbreaks. This is not to say that Microsoft products would not still be vulnerable to virus attacks, but at least they would not be drastically more vulnerable than the competition. Ironically, a disintegrated product line is more or less the fate that Microsoft appears to have avoided recently in its antitrust battle with the Justice Department. One day Microsoft may be seen as its own worst enemy.

That subtle differences in the mechanism of disease spread can be translated into distinct versions of the general percolation framework—possibly with quite different outcomes—suggests a certain amount of care is required in applying the methods of physics to the problem of epidemics. In the next chapter, in fact, we will see that other distinctions must be made if we are to understand the difference between biological contagion and social contagion problems like the diffusion of a technological innovation, distinctions that again carry important implications for the real-world phenomena we would like to understand. Percolation models, however, are so naturally applied to networks that they will continue to play an important role in the study of network epidemics. And as Mark and I soon realized, percolation is interesting for other reasons as well. Once again, however, it was László Barabási and Réka Albert who were one step ahead.

FAILURES AND ROBUSTNESS

LIKE MOST FEATURES OF COMPLEX SYSTEMS, GLOBAL CONNECTIVITY is neither unambiguously a good nor a bad thing. In the context of infectious diseases or computer viruses, the existence of a percolating

cluster in a network implies a potential epidemic. But in the context of a communication network like the Internet, where we would like to guarantee that packets of data will reach their destination in some reasonable time, then a percolating cluster would seem an absolute necessity. From the perspective of protecting infrastructure, therefore, from the Internet to airline networks, it is the *robustness* of the network's connectivity, with respect to accidental failures or deliberate attacks, that we want to preserve. And from this point of view also, percolation models can be extremely useful.

Having shown that a number of real networks like the Internet and the World Wide Web were what they called *scale-free*, Albert and Barabási started to wonder whether scale-free networks bore any comparative advantages over the more traditional varieties. Remember that in a scale-free network, the distribution of degree is governed by a power law instead of the sharply peaked Poisson distribution that we find in uniform random graphs—a distinction that, in practice, translates to a small fraction of “rich” nodes having very many links, and many other “poor” nodes having hardly any at all. Now Albert and Barabási became interested in the question of how well connected two networks, one a uniform random network and the other a scale-free network, would remain once their individual nodes started to fail.

Thinking about network robustness as a connectivity issue mapped the problem neatly into one about site percolation. In this application, however, the occupation probability played the opposite of its role in disease spreading. Whereas Mark and I had been primarily interested in the effect of occupied (susceptible) sites, Albert and Barabási were concerned with unoccupied sites—in network terms, the nodes that had failed. And in terms of robustness, the less effect that each unoccupied site had on the connectivity of the network, the better. Albert and Barabási also had a different view of connectivity from the one that Mark and I had used. Whereas we were concerned only with whether a percolating cluster existed or not, they wanted to know precisely how many steps would be required for a message to cross from

one side of the cluster to the other. Neither definition is universally the right way to think about robustness, but theirs was clearly relevant to systems like the Internet, where an increase in the typical number of hops taken by a message increases both its expected delivery time and its likelihood of being dropped.

The first thing that Albert and Barabási showed was that scale-free networks are far more resistant to *random* failures than are ordinary random networks. The reason is simply that the properties of scale-free networks tend to be dominated by the small fraction of highly connected *hub* nodes. Because they are so rare, these hubs are much less likely to fail by random chance than their less connected and far more plentiful counterparts. And like the absence of a minor rural airport from the airline network of the United States, the loss of a “poor” node goes largely unnoticed outside its immediate vicinity. In ordinary random networks, by contrast, the most-connected nodes are not nearly so critical, nor are the less well-connected nodes so inconsequential. As a result, every lost node will be missed—maybe not a great deal but more so than in a scale-free network. Invoking recent evidence that the Internet is in fact scale-free, Albert and Barabási went on to propose their model as an explanation of how the Internet works so reliably, even though individual routers fail all the time.

There is another side to robustness, however, that they also pointed out. Although in some networks like the Internet, router failures do occur at random, failures also can be a consequence of deliberate attacks, which may not be random at all. Even in the Internet, denial-of-service attacks, for example, tend to target highly connected nodes. And in other examples, from airline networks to communication networks, it is the hubs that are clearly the prime targets of any potential saboteur. What Albert and Barabási showed was that when the most highly connected nodes in a network are the first to fail, scale-free networks are actually much less robust than uniform networks. Ironically, the vulnerability of scale-free networks to attack is due to exactly the same property as their apparent robustness: in a

scale-free network, the most connected nodes are so much more critical to overall network functionality than their counterparts in a uniform network. The overall message, therefore, is an ambiguous one: the robustness of a network is highly dependent on the specific nature of the failures, with random and targeted failures offering diametrically opposite conclusions.

Although both kinds of failure are important to consider, the preferential failure of hubs seems particularly significant because it needn't be either deliberate or malicious. In many infrastructure networks that depend disproportionately on a small fraction of highly connected nodes, higher-than-average failure rates for those nodes may actually be an unavoidable consequence of their connectivity. For example, in the airline network, the massive amount of traffic passing through the major hubs increases their tendency to fail, a phenomenon with which air travelers in New York are painfully familiar. At LaGuardia Airport in Queens, both incoming and outgoing flights are stacked so close together that even a series of trivial delays, which at a small airport would be absorbed by the normal interval between flights, can accumulate to keep planes on the ground for hours, even on a picture-perfect day. In the year 2000, in fact, LaGuardia was the origin for 127 of the 129 most delayed flights in the country! And delays at hubs like LaGuardia are not just a problem for local travelers. Each flight delayed at a major hub tends to generate knock-on delays at its destination airport as well. So the more flights a hub handles, the greater its own chance of experiencing delays, and the greater the chance those delays will reverberate throughout the system.

The heavy dependence of modern airline networks on a subnetwork of hubs, therefore, causes them to be particularly susceptible to occasional widespread delays. But it also suggests a solution. Rather than persisting with a system in which the hubs bear all the burden of getting people from point A to point B, airlines could shift some of the links from the largest, most failure-prone hubs to the smaller regional airports whose delays derive principally from problems originating at

the hubs. Under such an arrangement, airports in Albuquerque and Syracuse, for example, would be connected directly, rather than having to route flights through Chicago or St. Louis. Very small airports, like those in Ithaca and Santa Fe, meanwhile would remain spokes. By reducing the effective degree of the hubs, the network as a whole would retain much of the efficiency that it derives from their large scale but would reduce the probability of individual failure. And even in the event that a hub did fail, fewer flights would be affected, causing the system as a whole to suffer less.

As straightforward as it seems in retrospect, Albert and Barabási's result was pretty neat, and with their paper on "Network attack and failure" gracing the cover of *Nature*, it generated a minor storm of media attention. We once again kicked ourselves for missing an obvious problem, and then, with the help of another of Steve's students—Duncan Callaway—scrambled to catch up. Duncan, in fact, succeeded in solving a much more difficult problem than the one Barabási's group had tackled. Using the techniques that Mark, Steve, and I had developed for studying the connectivity of random networks, Duncan managed to compute the different percolation transitions exactly, rather than just using computer simulations. He also managed to solve the problem for both link and node failures, and showed how to apply the model not just to scale-free networks but to random networks with any kind of degree distribution at all. All in all, it was an impressive effort, and the four of us managed to get a very nice paper out of it. But ultimately it didn't make a lot of difference. To a rough approximation, our findings were much the same as Albert and Barabási's, and we had to admit, they had thought of it first.

Fortunately for us, applying percolation techniques to the problems of the real world is a somewhat subtle business, so there were plenty of interesting problems left. Not only are real networks more complicated than any random model—scale-free or otherwise—but also the nature of the process itself is often poorly represented by the standard assumptions of percolation theory. Percolation models, for

example, typically assume that all nodes have the same likelihood of being susceptible. In reality, however, heterogeneity is an important feature of human and many nonhuman populations. Even in matters like disease spreading, individuals can vary widely in their inherent susceptibility or their capacity to infect others. And when behavioral and environmental factors are considered, the differences across a population can be complicated by the presence of strong correlations. It is often the case in sexually transmitted diseases, for example, that high-risk individuals are significantly more likely to interact with other high-risk individuals, a behavioral characteristic that may have social origins but clearly has epidemiological consequences.

Furthermore, the states of individuals can be correlated not only according to their intrinsic characteristics but also dynamically. A good analogy is the cascading failure in the power transmission grid discussed in chapter 1. If you were to assign failure probabilities to nodes at random, even taking account of their individual differences, you would still be missing an essential part of the problem: the role of *contingency*. The massive failure that occurred on August 10, 1996, remember, was not a result of multiple independent failures but rather was a *cascade* of failures, each one of which made subsequent failures more likely. Cascades of contingent, interdependent failures are more complicated to model than the percolation problems we have dealt with so far, but they happen all the time, and not just in engineering systems like the power grid. In fact, possibly the most widespread and interesting group of cascade problems lies in the realm of social and economic decision making. It is to these important, fascinating, and deeply mysterious problems that we now turn.