**Graphs and Networks** Lecture 6

Percolation I

*Daniel A. Spielman* September 17, 2013

## 6.1 Disclaimer

These notes are not necessarily an accurate representation of what happened in class. They are a combination of what I intended to say with what I think I said. They have not been carefully edited.

You should be able to find a diary of my Matlab session from today's class. It may reveal computations that do not appear in these notes.

## 6.2 Overview

We begin by introducing percolation and its connection to the SIR model of epidemics in networks. We then conduct a detailed study of percolation in infinite trees. Much of this study will be in terms of the Galton-Watson process.

In the next lecture, we study percolation in the grid and in the graphs from problem set 1.

## 6.3 Introduction

Percolation is essentially the study of what happens to a graph when one chooses to remove vertices or edges at random. In particular, one usually chooses some probability, $q$, and then chooses to remove each edge (or vertex) independently with probability $q$. We typically study the size and shape of the largest connected component in the remaining graph. We will only study the case in which edges are removed. This is what Physicists call *bond percolation.*

The study of percolation began with the study of a random medium in Physics. For example, they considered what happens when one puts a porous rock in a bucket of water. Will the water reach the center of the rock? The answer essentially depends on just how porous the rock is. Physicists model the rock as a graph in which water can flow along the edges. As the structure of the rock is somewhat random, they begin with a base graph, such as a 3-dimensional grid, and then close off edges with probability $q$. The water can reach the center of the rock if there is a path from the boundary of the graph to the inside.

Percolation also arises in the study of epidemics. While the book discusses many models of the spread of disease, I will just discuss the SIR (succeptable-infected-recovered) model. We would like to understand the chance that a disease spreads throughout a population after an initial person has been infected. We presume that each person has a chance to transmit the disease to the others in their social network[1]. A person can be in one of three states:

> Susceptible: they have not yet been exposed to the disease.
>
> Infected: they have the disease, and can pass it to others.
>
> Recovered: they had the disease, and are now immune (or removed).

Time clearly plays a role in understanding how the disease spreads, which is why you see so many differential equations in the percolation section of the book. But, we can analyze how far the disease spreads without reference to time.

Once a person becomes infected, we assume that there is some probability $p$ that they will transmit the disease to each other person in their social network. For simplicity, we will assume that these probabilities are independent, and the same for every pair of people in the network. After a person is finished being infected, they enter the recovered state. So, at the end of time, everyone is either susceptible or recovered.

A simple way to simulate this process is to begin with one infected node, which we call a *seed*. Every other node begins in state S. For every edge leaving the seed, with probability $p$ we set to infected the state of the vertex at the other end of that edge. We then set the state of the seed to recovered. While there remains an infected node, we repeat this process. Except, we do not allow nodes whose state is recovered to become infected again. The process stops when the state of every node is susceptible or recovered.

Alternatively, we can simulate this process without a notion of time by removing every edge from the network with probability $q = 1 - p$. If we then choose one “seed” person to infect, the set of people who eventually become recovered will be exactly those who are in the component of the graph containing the seed: the choices of the edges attached to the other people don’t matter.

The fundamental finding in studies of percolation in most artificial graphs is that it satisfies a threshold phenomenon. There is some probability $p_c$ so that if $p > p_c$, then the sampled network probably has a very large connected component. It is called the “giant component”, and usually contains a constant fraction of the vertices, with the exact number depending on how much $p$ exceeds $p_c$. On the other hand, if $p < p_c$, then all of the components in the graph are usually small.

The implication for vaccination is obvious: when you vaccinate someone, you remove them from the graph. So, if you are going to choose who to vaccinate, then you should vaccinate people so as to increase $p_c$ as much as possible.

Unfortunately, it is very difficult to understand percolation in an arbitrary graph: the best way to understand the process is often just by simulating it. However, we can understand how it behaves in some particular abstract graphs. We will prove the existence of thresholds in some of these, and supplement our understanding with experiments from problem set 1.

[1] We ignore for now the fact that some people are more likely to transmit a disease to you than others. I am much more likely to catch a cold from my kids than from my colleagues. So, we should really consider a weighted social network.

## 6.4 The Infinite Binary Tree and the Galton-Watson Process

We begin our analysis of percolation by considering percolation on the infinite binary tree. One way to define this is to identify the vertex set with strings over $\{0, 1\}$. The root is the empty string. Edges connect vertices with their immediate prefixes. So, 0 and 1 are the children of the root, 00 and 01 are the children of 0, etc.

We will keep each edge with probability $p$. We will prove that if $p > 1/2$, then there is a positive probability that the root is in an infinite component. Conversely, if $p < 1/2$, then the probability that the root is in an infinite component is zero.

Before entering the analysis, which we will do 3 different ways, I point out that this is equivalent to the Galton-Watson process. In the Galton-Watson process, one considers the progeny of a single-celled organism that reproduces by division. The probability $p$ is the chance that any one of the organisms survive. The first organism corresponds to the root of the tree. It splits into two children, which correspond to the vertices 0 and 1. Each survives to reproduce only if the edge connecting it to its parent appears in the tree. We will examine the probability that the descendants of the first cell continue to exist forever, or whether they die out. As you would expect, the threshold for this is exactly when the expected number of surviving progeny of each cell is 1.

So, the expected number of organisms in the first generation that survive to reproduce is $2p$. Similarly, the expected number of organisms in the second generation that survive to reproduce is $4p^2$. One way of seeing this is to identify 4 potential organisms in the second generation: the first and second child of each of the first and second children of the original. Each of these 4 potential organisms exists and survives only if their parent organism survives and they survive themselves. The probability of each of these events is $p^2$. We may similarly compute that the expected number of organisms in the $k$th generation is

$$2^k p^k = (2p)^k.$$

This is also the expected number of edges at the $k$th level of the tree that are connected to the root. For $p < 1/2$ this number goes to zero, whereas for $p > 1/2$ it goes to infinity. This is clearly some type of threshold phenomenon.

We can use this expectation calculation to prove it is unlikely that the descendants of the organism will survive for a long time when $p < 1/2$. Let $X^k$ be the random variable counting the number of descendants of the organism in the $k$th generation. The descendants are still around in the $k$th generation if and only if $X^k \geq 1$. However, Markov's inequality tells us that

$$\mathbf{Pr}\left[X^k \geq 1\right] \leq \mathbf{E}\left[X^k\right] \leq (2p)^k \xrightarrow[k\to\infty]{} 0$$

But, what about when $p > 1/2$? The expected number of descendants goes to infinity. But, what does that tell us about the chance that the number of descendants is in fact infinite? This is less obvious.

## 6.5 $p > 1/2$

Let $\theta_p$ be the probability that the root of the tree is in an infinite component. We will prove that for $p > 1/2$, $\theta_p > 0$.

For each node $a$, let $E_a$ be the event that $a$ is the root of an infinite component (that is, that a path from $a$ going forward extends forever). By definition

$$\theta_p = \mathbf{Pr}\left[E_\emptyset\right].$$

As the infinite binary tree is self-similar,

$$\mathbf{Pr}\left[E_a\right] = \theta_p$$

for all nodes $a$. We also know that $E_\emptyset$ happens only if the edge between $\emptyset$ and $a$ appears in the graph for $a \in \{0, 1\}$, and if node $a$ is the root of an infinite component. Let $F_a$ be the event that the edge between $\emptyset$ and $a$ appears in the graph and $E_a$ holds. Then,

$$E_\emptyset = F_0 \text{ or } F_1.$$

So,

$$\mathbf{Pr}\left[E_\emptyset\right] = \mathbf{Pr}\left[F_0\right] + \mathbf{Pr}\left[F_1\right] - \mathbf{Pr}\left[F_0 \text{ and } F_1\right].$$

As $F_0$ and $F_1$ are independent even,

$$\mathbf{Pr}\left[E_\emptyset\right] = \mathbf{Pr}\left[F_0\right] + \mathbf{Pr}\left[F_1\right] - \mathbf{Pr}\left[F_0\right]\mathbf{Pr}\left[F_1\right].$$

Observing that

$$\mathbf{Pr}\left[F_a\right] = p\,\mathbf{Pr}\left[E_a\right] = p\theta_p,$$

we obtain

$$\theta_p = \mathbf{Pr}\left[E_\emptyset\right] = 2p\theta_p - p^2\theta_p^2.$$

This gives us an equation for $\theta_p$. One solution is $\theta_p = 0$. When $\theta_p \neq 0$, we can divide by $\theta_p$ to get

$$2p - 1 = p^2\theta_p, \qquad \Longrightarrow \qquad \theta_p = \frac{2p-1}{p^2}.$$

For $p > 1/2$, this is a probability greater than 0. As $p$ approaches 1, $\theta_p$ does too. For $p < 1/2$, $\theta_p$ is negative. As we cannot have a negative probability, the solution must be 0 in this case.

## 6.6 Finite Binary Trees

In case the analysis of infinite trees makes you uncomfortable, we will perform an analysis of finite binary trees of increasing depth. As before, let $p$ be the probability that each edge is chosen to appear in the graph. We define $\theta_{p,k}$ to be the probability that the root is connected to a leaf in a tree of depth $k$, where the tree consisting of only the root has depth 0. In particular, $\theta_{p,0} = 1$.

This is the same as the probability that the root has an edge to a child that connects to a leaf in a tree of depth $k-1$. As before, let $E_a$ be the event that node $a$ is connected by a forwards path to a leaf. We can again compute

$$\theta_{p,k} \overset{\text{def}}{=} \mathbf{Pr}\left[E_\emptyset\right] = p\,\mathbf{Pr}\left[E_0\right] + p\,\mathbf{Pr}\left[E_1\right] - p^2\,\mathbf{Pr}\left[E_0\right]\mathbf{Pr}\left[E_1\right] = 2p\theta_{p,k-1} - p^2\theta_{p,k-1}^2.$$

This gives us a recurrence for $\theta_{p,k}$. Since $\theta_{p,0} = 1$, we derive

$$\theta_{p,1} = 2p - p^2.$$

As $k$ grows large, we expect $\theta_{p,k}$ to approach a limit. If it does, it should be a number $q$ that satisfies the equation

$$q = 2pq - (pq)^2.$$

Our analysis from the previous section showed that $q = 0$ is a solution, and that for $p > 1/2$ there is another solution:

$$q \overset{\text{def}}{=} \frac{2p-1}{p^2}.$$

We will now show by induction that

$$\theta_{p,k} \geq q$$

for all $k \geq 0$. To see this, we examine the function

$$f(x) = 2px - (px)^2,$$

as

$$\theta_{p,k} = f(\theta_{p,k-1}).$$

We will base our induction in the case $k = 0$, for which we have

$$1 = \theta_{p,0} \geq q$$

To perform the induction, we will show that $f(x) \geq q$ for $x \in [q, 1]$. We first compute the derivative of $f$ with respect to $x$ and find

$$f'(x) = 2p - 2p^2x = 2p(1 - px) > 0$$

for $x \in (0, 1]$. This means that $f$ is an increasing function on $(0, 1]$. As $f(q) = q$, we may conclude that $f(x) \geq q$ for $x \geq q$. Thus,

$$\theta_{p,k} \geq q$$

for all $k \geq 0$.

With a little more work one can show that

$$\liminf_{k\to\infty} \theta_{p,k} = q.$$

One consequence of this is that with probability at least $q$ the descendants of the organism never die out. That is, they exist for an infinite number of generations.

## 6.7 The Number of Descendants and $k$-ary trees

We will now do a more detailed analysis in which we examine the number of descendants of an organism. We will perform this analysis in a more general setting: Each organism will divide into $k$ others. That is, we will do percolation on the infinite $k$-ary tree. This is what the Physicists call the Bethe Lattice.

We set the probability that an organism survives to reproduce to $p = c/k$. In the sub-critical regime ($c < 1$) we will see that it is very unlikely that the organism has too many descendants. In the super-critical regime ($c > 1$) we will see that once the number of descendants of an organism becomes sufficiently large it is likely to be infinite. Just as in the case with $k = 2$, we can prove that for $c > 1$ there is a constant probability of an organism spawning an infinite number of generations.

We will find it useful to assign a number of every cell that survives to reproduce. We number the first cell 1. We must use consecutive numbers in a consistent manner, and must assign every cell a lower number than each of its descendants. For example, if there are $j$ cells in the first generation that survive to reproduce, we could assign them numbers 2 through $j + 1$. We could then assign numbers to the cells in the second generation, and so on.

For each $j$ such that cell $j$ survives to reproduce, we introduce Bernoulli random variables $X_{j,1}, \dots, X_{j,k}$ where $X_{j,i} = 1$ if the $i$th child of cell $j$ survives to reproduce. So, the number of descendants of cells 1 through $u$ is

$$1 + \sum_{j=1}^{u} \sum_{i=1}^{k} X_{j,i} - u.$$

Cell $u$ is the last surviving member of the population precisely when

$$1 + \sum_{j=1}^{u} \sum_{i=1}^{k} X_{j,i} = u$$

and for all $v < u$

$$1 + \sum_{j=1}^{v} \sum_{i=1}^{k} X_{j,i} > v.$$

We will now use the Chernoff bounds to bound how unlikely this is in the sub-critical case. Define

$$X^{(u)} = \sum_{j=1}^{u} \sum_{i=1}^{k} X_{j,i}.$$

The expectation of $X^{(u)}$ is

$$\mu = ukp = uk\frac{c}{k} = uc.$$

For $c < 1$ this becomes significantly less than $u$ and the Chernoff bounds will imply that $X^{(u)}$ is very unlikely to be more than $u$. Before we carry out the details of that argument, let me put one issue to rest. You might worry that $X^{(u)}$ is only defined if cell $u$ actually survives to reproduce. You may then worry about what it means to take this sum if $X^{(u-1)} < u-1$. To make these notions precise, consider sampling all the variables $X_{j,i}$ for $1 \leq j \leq u$ and $1 \leq j \leq k$ without thinking about the Galton-Watson process. If it turns out that organism $j$ does survive to reproduce, then and only then look at the variables $X_{j,i}$ to figure out which of its children survive to reproduce. If organism $j$ never exists, then just throw away the unused variables[2].

Let $Z$ be the number of descendants of the first organism, plus 1 for the first organism (or view 1 as a descendant of itself). We can now say that

$$\mathbf{Pr}\left[Z > u\right] \leq \mathbf{Pr}\left[X^{(u)} \geq u\right] \leq \exp\left(-\frac{1}{3}\delta^2\mu\right),$$

where we set $\delta$ so that

$$\begin{aligned}(1+\delta)\mu &= u\\ (1+\delta)uc &= u\\ (1+\delta) &= \frac{1}{c}\\ \delta &= \frac{1}{c} - 1,\end{aligned}$$

which is greater than 0 in the sub-critical case. We conclude that

$$\mathbf{Pr}\left[Z > u\right] \leq \exp\left(-\frac{1}{3}\frac{(1-c)^2}{c}u\right).$$

So, the probability that there are more than $u$ descendants decreases exponentially with $u$.

In the super-critical case we will perform a similar analysis. We will show that it is very unlikely that $Z = u$ for any sufficiently larger $u$. By summing over all large $u$ we will conclude that if $Z$ is not small then it is probably infinite. Here the expectation of $X^{(u)}$ is also $cu$, but $c > 1$. We have

$$\mathbf{Pr}\left[Z = u\right] \leq \mathbf{Pr}\left[X^{(u)} \leq u\right] \leq \exp\left(-\frac{1}{2}\delta^2\mu\right),$$

where we set $\delta$ so that

$$\begin{aligned}(1-\delta)cu &= u\\ (1-\delta) &= \frac{1}{c}\\ \delta &= 1 - \frac{1}{c},\end{aligned}$$

which is greater than zero in the super-critical case. We thereby conclude that

$$\mathbf{Pr}\left[Z = u\right] \leq \exp\left(-\frac{1}{2}\frac{(c-1)^2}{c}u\right) = \exp\left(-\frac{1}{2}\frac{(c-1)^2}{c}\right)^u.$$

Define

$$\gamma = \exp\left(\frac{1}{2}\frac{(c-1)^2}{c}\right).$$

[2] This may worry you, but I assure you that you can make it formal.

By summing an infinite series we can now bound the probability that $Z$ is a large but finite number. We have

$$\mathbf{Pr}\left[u \leq Z < \infty\right] = \sum_{w=u}^{\infty} \mathbf{Pr}\left[Z = w\right] \leq \sum_{w=u}^{\infty} \gamma^{-w} = \frac{\gamma^{-u}}{1-\gamma^{-1}}.$$

So, this probability also decreases exponentially with $u$. This tells us that once we know a cell has a moderate number of descendants, it becomes very unlikely that its progeny die out. Another way of saying this is that its descendants are probably few or infinite.