An Infection Spreading Model on Binary Trees

Daniel Guo

Mentored by Professor Partha Dey

MIT PRIMES-USA

Abstract

An important and ongoing topic of research is the study of infectious diseases and the speed at which these diseases spread. Modeling the spread and growth of such diseases leads to a more precise understanding of the phenomenon and accurate predictions of spread in real life. We consider a long-range infection model on an infinite regular binary tree. Given a spreading coefficient $\alpha > 1$, the time it takes for the infection to travel from one node to another node below it is exponentially distributed with specific rate functions such as $2^{-k}k^{-\alpha}$ or $\frac{1}{\alpha^k}$, where k is the difference in layer number between the two nodes. We simulate and analyze the time needed for the infection to reach layer m or below starting from the root node. The resulting time is recorded and graphed for different values of α and m. Finally, we prove rigorous lower and upper bounds for the infection time, both of which are approximately logarithmic with respect to m. The same techniques and results are valid for other regular d-ary trees, in which each node has exactly d children where $d > 2$.

Contents

1 Introduction **1**

2 Methods and Techniques **2**

2.1 The Model 2

2.2 Algorithm and Simulation 4

2.3 Proof Techniques for Lower and Upper Bounds 4

3 Proof of Lower and Upper Bounds **5**

4 Illustrations **16**

5 New Rates **17**

6 Discussion **21**

7 Conclusions and Future Work **22**

8 Acknowledgements **23**

1 Introduction

The spread of infection over certain data structures has been studied and modeled extensively. The growth pattern and time of spreading over a one-dimensional percolation model are investigated in [1] and [2]. The results established in these papers are extended to the d-dimensional lattice $\mathbb{Z}^d$ in [3]. In [3], infection times are independent exponential distributions with rates that are dependent on the distance between lattice points.

However, there are still a number of other areas that have not been researched, particularly pertaining to variations in graphical data structures and distributions of infection time. This project answers some of these problems by extending the investigation to a structure with homogeneity and ordering, namely a perfect directed binary tree. A binary tree is chosen because it is the simplest symmetrical tree to operate on, in comparison with ternary trees and other n-ary trees. However, the same methods and results can be applied to any regular n-ary trees. In addition, the condition of perfection, or the requirement that each node has exactly two child nodes, is used in order to create homogeneity: at layer m, there are 2^m nodes. For the spreading time of the disease, an exponential distribution with a rate based on distance is also chosen, in order to derive a result that is an extension of [3].

The problem is about the spread of an infection on a binary tree. The nodes are the "people", or in a more general term, the hosts of the disease. The infection starts at the root node and jumps from any initial node to any other node in a binary tree with the initial node as the root. Thus, each node has a chance to be infected at any given step, with nodes that are farther away having less of a risk. As stated above, the time it takes to infect between two nodes is given by an exponential distribution with rate as a function of the distance. With this model, we propose a general question to be answered: Given a previously-defined coefficient of expansion α (which affects the rate of the exponential distribution) and a destination layer m, what are some strong upper and lower bounds for the amount of time it takes for the infection to reach layer m or below, given that m is very large?

The spread of certain objects or ideas, such as an infectious disease or an interesting fact, is an important topic in today's world. The rapid increase in usage of the Internet is causing an increase of information transfer. It is possible to model this transfer with the stated binary tree model, treating information as an infection.

In addition, with the growing concern over various potent diseases, the investigation of the spread of diseases is of paramount importance. The mathematical model addresses this question. Indeed, by finding the approximate time of infection, we will predict the actual time in reality and take action to quarantine the disease or slow the disease's spread.

Because the number of nodes on each layer of the tree increases exponentially, we hypothesize that the lower and upper bounds on time would both be of the form $c_1 \ln(m) + c_2$, where c_1 and c_2 are arbitrary constants.

2 Methods and Techniques

2.1 The Model

Consider the infinite directed binary tree $\mathcal{T}$. Using σ to denote any node of $\mathcal{T}$, we define $|\sigma|$ as the layer that σ is on, which is conventional notation by [6]. The definition of the layer of σ is the number of edges in the shortest path from the root node $\emptyset$ to σ. As a result, the root node is considered to be on layer 0. We also consider the rate function

$$r(k) := \frac{1}{2^k k^\alpha} \tag{1}$$

where $\alpha > 1$ is a pre-determined constant and $k \geqslant 1$ is the difference in layer number between the two nodes. (We will investigate an alternative rate function later on in section 5.)The infection model is given by a random weighted graph, where each node is in $\mathcal{T}$. The edges of $\mathcal{T}$ are $\langle \sigma_0, \sigma_1 \rangle$ where σ_0, σ_1 are arbitrary nodes and σ_1 is in the sub-tree of σ_0. This definition generalizes the conventional edges of a binary tree. The time $\omega_{\sigma_0,\sigma_1}$ to cross the edge $\langle \sigma_0, \sigma_1 \rangle$ is an exponentially distributed random variable with rate $r(|\sigma_1| - |\sigma_0|)$ and is independent for all edges. Because the mean of an exponential random variable is the reciprocal of its rate, a higher α means that the infection will spread slower.

The rate function in equation (1) is specifically chosen to simplify calculations. For any fixed node σ_0, there are exactly $2^{m-|\sigma_0|}$ nodes that are on layer m and in the subtree of σ_0 for $m > |\sigma_0|$. Iterating m from $|\sigma_0|$ to ∞, we know that the sum of the rates for the stated nodes will be $\sum_{m>|\sigma_0|}^{\infty} 2^{m-|\sigma_0|} r(m - |\sigma_0|) = \sum_{m>|\sigma_0|}^{\infty} \frac{1}{(m-|\sigma_0|)^\alpha}$. The finiteness of the total rate is necessary for the model to be well-defined. Otherwise, instantaneous infection will occur, as shown below.

At time 0 only the root $\emptyset$ is infected. Given a path $P = (\sigma_0, \sigma_1, \ldots, \sigma_k)$, with the requirement that σ_{i+1} is in the sub-tree of σ_i as stated above, the time associated with the path is $T(P) = \sum_{i=1}^{k} \omega_{\sigma_{i-1},\sigma_i}$. The time to reach a node $\sigma \neq \emptyset$ is given by

$$T(\emptyset, \sigma) := \inf_{P \in \mathcal{P}} T(P)$$

where $\mathcal{P}_\sigma$ is the collection of all paths P such that $\sigma_0 = \emptyset$, $\inf$ denotes infimum, and $\sigma_k = \sigma$. The time to reach layer m or below is given by

$$T_m := \inf_{\sigma : |\sigma| \geqslant m} T(\emptyset, \sigma).$$

We want to understand the growth rate and limiting behavior of T_m as $m \to \infty$ for different values of α. We do so by bounding T_m as a function of m.

Note that

$$\sum_{\sigma \in \mathcal{T} \setminus \emptyset} r(|\sigma|) = \sum_{k=1}^{\infty} 2^k r(k) = \sum_{k=1}^{\infty} \frac{1}{k^\alpha} < \infty$$

only when $\alpha > 1$. Thus we will take $\alpha > 1$, to ensure non-instantaneous growth of infection.

2.2 Algorithm and Simulation

A simulation is used to give an approximation of the bounds. The algorithm is implemented using MATLAB. The binary tree is given as a vector, with each node index pointing to the node directly above it. Because each node has a probability to be infected based on an exponential random distribution, we use the properties of such a distribution to make our work easier. The algorithm works as follows: at each step, an infected node σ_{inf} is randomly chosen. From this, a time is generated from an exponential random variable with a rate equal to the total rate of each node that can be infected from σ_{inf}. After this is done, a new node, the one that will be infected next, is selected. If the infected node is at layer m or below, the program stops. In the program, a slightly different rate function of $r_0(k) = 2^{-k}((k^{1-\alpha} - (k+1)^{1-\alpha})$ is used to make the simulation faster. The ratio of $\frac{r_0(k)}{r(k)}$ approaches $\alpha - 1$ as k grows large for any α.

2.3 Proof Techniques for Lower and Upper Bounds

By finding the moment generating function (MGF) of an exponential random variable and using Markov's Inequality, we bound the probability that the infection time T_m is lower than a certain time. Given any path P, we write $T(P)$ as a sum of exponential random variables with arbitrary rates. Since the time $T(\emptyset, \sigma)$ is the infimum of all possible paths $P \in \mathcal{P}$, and each $T(P)$ is the sum of exponential random variables with arbitrary rates, we use union bounds and concentration properties for $T(P)$ to get a lower bound for $T(\emptyset, \sigma)$. For the upper bound on time, we prove that the time it takes for all nodes at layer k to be infected is ck, and then bound the time that it takes from the infection to travel from layer k to layer m, also using Markov's Inequality.

There are five properties of an exponential random variable that are used to understand the distribution and help prove the bounds.

Lemma 2.1. *We have the following:*

(a) *The cumulative distribution function* (CDF) *of an exponential distribution with rate* r *is* $1 - e^{-rx}, x > 0$.

(b) *If $X \sim$ Exp(r), where $Exp(r)$ denotes an exponential random variable with rate r, then*

$$\mathbb{P}(X > x + y | X > y) = \mathbb{P}(X > x) \text{ for any } x, y > 0.$$

(c) *(Minimum Property) If $X_1, X_2, \ldots, X_n$, are independent exponential random variables with rates $r_1, r_2, \ldots, r_n$, respectively, then $X = \min(X_1, X_2, \ldots, X_n)$ is an exponential random variable with rate $r_1 + r_2 + \cdots + r_n$.*

(d) *(Location of the Minimum) If $X_1, X_2, \ldots, X_n$, are independent exponential random variables with rates $r_1, r_2, \ldots, r_n$, respectively, then the probability that X_i is the minimum of $X_1, X_2, \ldots, X_n$ is $r_i/(r_1 + r_2 + \cdots + r_n)$. In other words,*

$$\mathbb{P}(X_i = \min(X_1, X_2, \ldots, X_n)) = \frac{r_i}{r_1 + r_2 + \cdots + r_n}.$$

(e) *If X is an exponential random variable with rate 1, then $\frac{X}{r}$ is an exponential random variable with rate r.*

Proofs of these properties are found in [5].

3 Proof of Lower and Upper Bounds

Consider the infection model on a binary tree defined in section 2.1 with rate function $r(k) = 2^{-k}k^{-\alpha}$ for some $\alpha > 1$. Define T_m as the time needed for the infection to reach layer m or below starting from the root node $\emptyset$.

Theorem 3.1. *(Lower Bound) For any $\varepsilon > 0$, we have*

$$\mathbb{P}(T_m \geq \frac{\alpha - 1}{c}\ln(m-1) - \frac{1}{c}\ln((c\varepsilon)^{-1}e\ln(m-1))) \geq 1 - \varepsilon$$

where $c = 2^\alpha\zeta(\alpha)$ and $\zeta(\alpha)$ is the Riemann zeta function.

Theorem 3.2. *(Upper Bound) Using the binary tree spreading method defined in section 2.1, the procedures described in section 2.3, and T_m as defined in theorem 3.1, we have that*

$$\mathbb{P}(T_m \leq \frac{c(\alpha-1)}{\ln 2}\ln m + c\log_2 \frac{e(\alpha-1)\ln 2}{c}) > 1 - ((2ce^{1-c})^k + e^{-cr_{total}/\ln 2})$$

where $c > 3$, $k = \log_2((\alpha-1)\ln 2 \cdot m^{\alpha-1}) - \log_2 c$, and $r_{total} = \frac{2^k(m-k-1)^{-\alpha+1}}{\alpha-1}$.

First, we establish a few well-known results.

Definition 3.3 (Moment Generating Function)**.** *The* MGF *of a random variable X is defined as $M_X(t) = \mathbb{E}[e^{tX}]$, where $\mathbb{E}$ denotes the expected value.*

This is a widely-known definition and is found in [5].

Lemma 3.4. *The* MGF *of an exponential random variable X with rate r is $M_X(t) = \frac{r}{r-t}$ for $t < r$.*

Proof. We have

$$M_X(t) = \mathbb{E}[e^{tX}] = \int_{-\infty}^{\infty} e^{tx} \cdot re^{-rx}dx = r\int_0^{\infty} e^{(t-r)x}dx = r\left[\frac{1}{t-r}e^{(t-r)x}\right]\Bigg|_0^{\infty} = \frac{r}{r-t}.$$

Note that $t < r$, otherwise the integral evaluates to infinity. ■

Lemma 3.5. *The* MGF *of the sum of two random variables X and Y is $M_{X+Y}(t) = M_X(t) \cdot M_Y(t)$.*

A proof of this lemma is found in [5].

Lemma 3.6. *(Markov's Inequality) For any random variable W that takes on non-negative values, we have $\mathbb{P}(W > n) \leq \mathbb{E}(W)/n$.*

A proof of this lemma is found in [5].

Corollary 3.7. *For any random variable Y that takes on non-negative values, we have that $\mathbb{P}(Y < t) \leq e^{st}M_Y(-s)$ and $\mathbb{P}(Y > t) \leq e^{-st}M_Y(s)$.*

Proof. By letting $W = e^{\mp sY}$ and $n = e^{\mp st}$ in lemma 3.6, we get

$$\mathbb{P}(Y < t) = \mathbb{P}(-sY > -st) = \mathbb{P}(e^{-sY} > e^{-st}) \leq e^{st} M_Y(-s)$$
$$\mathbb{P}(Y > t) = \mathbb{P}(sY > st) = \mathbb{P}(e^{sY} > e^{st}) \leq e^{-st} M_Y(s)$$

under the condition that $t > 0$. ■

Lemma 3.8. *For any positive integer ℓ, we have*

$$e^{-\ell+1}\ell^{\ell+\frac{1}{2}} \geq \ell!.$$

Proof. The inequality follows trivially for $\ell = 1$. For $\ell \geq 2$, we have

$$\sqrt{2\pi}e^{\frac{1}{12\ell}-1} \leq \sqrt{2\pi}e^{\frac{1}{24}-1} \approx 0.961 \leq 1$$

and thus

$$\sqrt{2\pi}e^{\frac{1}{12\ell}-\ell}\ell^{\ell+\frac{1}{2}} \leq e^{-\ell+1}\ell^{\ell+\frac{1}{2}}.$$

Thus, the inequality is reduced to $\ell! \leq \sqrt{2\pi}e^{\frac{1}{12\ell}-\ell}\ell^{\ell+\frac{1}{2}}$ for all $\ell \geq 2$ (see [4] for a proof). ■

Lemma 3.9. *Let $X_1, X_2, \ldots, X_n$ be independent and identically distributed (i.i.d.) exponential random variables with rate* 1. *Define $Y := \frac{X_1}{r_1} + \frac{X_2}{r_2} + \cdots + \frac{X_n}{r_n}$. Then we have that*

$$\mathbb{P}\Big(\frac{X_1}{r_1} + \frac{X_2}{r_2} + \cdots + \frac{X_n}{r_n} \leq t\Big) \leq \Big(\frac{et}{n}\Big)^n \prod_{i=1}^{n} r_i.$$

Proof. Using corollary 3.7, lemma 2.1(e) and lemma 3.4, we have that

$$\begin{aligned}\mathbb{P}(Y < t) \leq e^{st}M_Y(-s) &= e^{st}\prod_{i=1}^{n} M_{\frac{X_i}{r_i}}(-s)\\ &= e^{st}\prod_{i=1}^{n}\frac{r_i}{r_i+s} \leq e^{st}\prod_{i=1}^{n}\frac{r_i}{s} = e^{st}s^{-n}\prod_{i=1}^{n} r_i.\end{aligned}$$

Since s can take on any real number greater than 0, the quantity above needs to be minimized as to make the inequality the strongest. Hence, we need to solve

$$\frac{d}{ds}(e^{st}s^{-n}) = e^{st}(ts^{-n} - ns^{-n-1}) = 0 \implies s = \frac{n}{t}.$$

Thus, we have

$$\mathbb{P}\Big(\frac{X_1}{r_1} + \frac{X_2}{r_2} + \cdots + \frac{X_n}{r_n} \leq t\Big) \leq \Big(\frac{et}{n}\Big)^n \prod_{i=1}^{n} r_i,$$

which was to be proved. ■

Lemma 3.10. *Let $X_1, X_2, \ldots, X_k$ be i.i.d. exponential random variables with rate 1. Then, given that $k < t$,*

$$\mathbb{P}(X_1 + X_2 + \cdots + X_k \geq t) \leq e^{-t}\Big(\frac{et}{k}\Big)^k.$$

Proof. Using corollary 3.7, lemma 2.1(e) and lemma 3.4, we have that

$$\mathbb{P}(X_1 + X_2 + \cdots + X_k \geq t) \leq e^{-st}M_{X_1+X_2+\cdots+X_k}(s) = e^{-st}\prod_{i=1}^{k}\frac{1}{1-s} = e^{-st}(1-s)^{-k}.$$

As with the proof of the lower bound, we find the value of s to minimize the right-hand side of the inequality, which is equivalent to setting the derivative with respect to s to 0:

$$\frac{d}{ds}(e^{-st}(1-s)^{-k}) = e^{-st}(-t(1-s)^{-k} + k(1-s)^{-k-1}) = 0 \implies s = 1 - \frac{k}{t}.$$

Substituting s, we have that

$$\mathbb{P}(X_1 + X_2 + \cdots + X_k \geq t) \leq e^{k-t}\Big(\frac{k}{t}\Big)^{-k} = e^{-t}\Big(\frac{et}{k}\Big)^k$$

which was to be proved. ■

Lemma 3.11. *(Union Bound) Given any events $A_1, A_2, \dots, A_i$, we have that*

$$\mathbb{P}\Big(\bigcup_i A_i\Big) \leq \sum_i \mathbb{P}(A_i).$$

A proof of this lemma is found in [5].

With these facts, we begin proving the main theorems.

Proof of Theorem 3.1. From now on, define $\sigma_0 = \emptyset$ and $\sigma_i < \sigma_j$ to mean that $|\sigma_i| < |\sigma_j|$ (recall that $|\sigma|$ denotes the layer of σ) and σ_j is in the sub-tree of σ_i. This is conventional notation by [6]. Note that

$$\begin{aligned}\mathbb{P}(T_m \leq t) &= \mathbb{P}\Big(\inf_{\sigma:|\sigma|\geqslant m} T(\emptyset,\sigma) \leq t\Big) = \mathbb{P}\Big(\bigcup_{\sigma:|\sigma|\geqslant m} T(\emptyset,\sigma) \leq t\Big) \\ &\leq \sum_{\sigma:|\sigma|\geqslant m} \mathbb{P}(T(\emptyset,\sigma) \leq t) = \sum_{\sigma:|\sigma|\geqslant m} \mathbb{P}\Big(\inf_{P\in\mathcal{P}_\sigma} T(P) \leq t\Big).\end{aligned}$$

The inequality is derived from lemma 3.11, where each A_i is equivalent to $T(\emptyset,\sigma) \leq t$. In addition, because each path in $\mathcal{P}_\sigma$ can be described as a set of nodes $(\sigma_0, \sigma_1, \dots, \sigma_k)$ with $\sigma_0 = \emptyset$ and $\sigma_k = \sigma$, we get that

$$\begin{aligned}\sum_{\sigma:|\sigma|\geqslant m} \mathbb{P}\Big(\inf_{P\in\mathcal{P}_\sigma} T(P) \leq t\Big) &= \sum_{\sigma:|\sigma|\geqslant m} \mathbb{P}\Big(\bigcup_{P\in\mathcal{P}_\sigma} T(P) \leq t\Big) \leq \sum_{\sigma:|\sigma|\geqslant m} \sum_{P\in\mathcal{P}_\sigma} \mathbb{P}(T(P) \leq t) \\ &= \sum_{k=m}^{\infty} \sum_{\ell=1}^{m} \sum_{\sigma_1<\sigma_2<\cdots<\sigma_\ell:|\sigma_\ell|=k,|\sigma_{\ell-1}|<m} \mathbb{P}(T(P) \leq t) \\ &\leq \sum_{k=m}^{\infty} \sum_{\ell=1}^{m} \sum_{\sigma_1<\sigma_2<\cdots<\sigma_\ell:|\sigma_\ell|=k,|\sigma_{\ell-1}|<m} \Big(\frac{et}{\ell}\Big)^{\ell} \prod_{i=1}^{\ell} r(|\sigma_i| - |\sigma_{i-1}|).\end{aligned}$$

The first inequality follows from lemma 3.11, while the second inequality follows from lemma 3.9. Thus, we have to find an upper bound for

$$\sum_{k=m}^{\infty} \sum_{\ell=1}^{m} \sum_{\sigma_1<\sigma_2<\cdots<\sigma_\ell:|\sigma_\ell|=k,|\sigma_{\ell-1}|<m} \Big(\frac{et}{\ell}\Big)^{\ell} \prod_{i=1}^{\ell} r(|\sigma_i| - |\sigma_{i-1}|).$$

If we choose $t = t_m$ as a function of m such that the upper bound of this quantity is converging to 0 as $m \to \infty$, then $T_m \geq t_m$ with high probability.

Define $s(n) := 2^n r(n), n \geq 1$ and

$$f_\ell(n) := \sum_{\sigma_1 < \sigma_2 < \cdots < \sigma_\ell : |\sigma_\ell| = n} \prod_{i=1}^{\ell} r(|\sigma_i| - |\sigma_{i-1}|). \tag{2}$$

Notice that definition (2) requires that $n \geq \ell$, because if $n < \ell$, then ℓ nodes would be on $\ell - 1$ or fewer layers, contradiction. Moreover, we have

$$\begin{aligned}
f_\ell(n) &= \sum_{k=\ell-1}^{n-1} \sum_{\sigma_1 < \sigma_2 < \cdots < \sigma_{\ell-1} : |\sigma_{\ell-1}| = k} \prod_{i=1}^{\ell-1} r(|\sigma_i| - |\sigma_{i-1}|) \sum_{\sigma_{\ell-1} < \sigma_\ell} r(|\sigma_\ell| - |\sigma_{\ell-1}|) \\
&= \sum_{k=\ell-1}^{n-1} \sum_{\sigma_1 < \sigma_2 < \cdots < \sigma_{\ell-1} : |\sigma_{\ell-1}| = k} \prod_{i=1}^{\ell-1} r(|\sigma_i| - |\sigma_{i-1}|) \cdot 2^{n-k} \cdot r(n-k) \\
&= \sum_{k=\ell-1}^{n-1} s(n-k) \sum_{\sigma_1 < \sigma_2 < \cdots < \sigma_{\ell-1} : |\sigma_{\ell-1}| = k} \prod_{i=1}^{\ell-1} r(|\sigma_i| - |\sigma_{i-1}|) \\
&= \sum_{k=\ell-1}^{n-1} s(n-k) f_{\ell-1}(k).
\end{aligned}$$

We have to bound

$$\mathbb{P}(T_m \leq t) \leq \sum_{\ell=1}^{m} \left(\frac{et}{\ell}\right)^{\ell} \sum_{k=m}^{\infty} f_\ell(k).$$

We recall that $s(n) = 2^n r(n) \leq n^{-\alpha}$ for $k \geq 1$ and thus, $f_1(n) = s(n) \leq n^{-\alpha}$. Now, we will prove that

$$f_\ell(n) \leq 2^{\alpha(\ell-1)} \zeta(\alpha)^{\ell-1} n^{-\alpha} \text{ for all } \ell \geq 1 \tag{3}$$

by induction. We have already shown this for $\ell = 1$. Suppose the inequality (3) holds for all

$\ell \leq p-1$. For $\ell = p$, we have

$$\begin{aligned} f_p(n) &= \sum_{k=p-1}^{n-1} s(n-k) f_{p-1}(k) \\ &\leq \sum_{k=p-1}^{n-1} (n-k)^{-\alpha} \cdot 2^{\alpha(p-2)} \zeta(\alpha)^{p-2} k^{-\alpha} \\ &= 2^{\alpha(p-2)} \zeta(\alpha)^{p-2} \sum_{k=p-1}^{n-1} k^{-\alpha}(n-k)^{-\alpha}. \end{aligned}$$

Now,

$$\sum_{k=\ell-1}^{n-1} k^{-\alpha}(n-k)^{-\alpha} = n^{-\alpha} \sum_{k=\ell-1}^{n-1} \left(\frac{1}{k} + \frac{1}{n-k} \right)^{\alpha}.$$

Using the fact that $(x+y)^\alpha \leq 2^{\alpha-1}(x^\alpha + y^\alpha)$ for any $\alpha \geq 1, x, y > 0$, we have

$$n^{-\alpha} \sum_{k=\ell-1}^{n-1} \left(\frac{1}{k} + \frac{1}{n-k} \right)^{\alpha} \leq n^{-\alpha} \sum_{k=\ell-1}^{n-1} 2^{\alpha-1} \left(\frac{1}{k^\alpha} + \frac{1}{(n-k)^\alpha} \right) \leq n^{-\alpha} 2^\alpha \zeta(\alpha).$$

Combining, we have

$$f_p(n) \leq 2^{\alpha(p-2)} \zeta(\alpha)^{p-2} \sum_{k=p-1}^{n-1} k^{-\alpha}(n-k)^{-\alpha} \leq 2^{\alpha(p-1)} \zeta(\alpha)^{p-1} n^{-\alpha},$$

proving inequality (3).

Using inequality (3), we have

$$\begin{aligned} \mathbb{P}(T_m \leq t) \leq \sum_{\ell=1}^{m} \left(\frac{et}{\ell} \right)^{\ell} \sum_{k=m}^{\infty} f_\ell(k) &\leq \sum_{\ell=1}^{m} \left(\frac{et}{\ell} \right)^{\ell} \sum_{k=m}^{\infty} 2^{\alpha(\ell-1)} \zeta(\alpha)^{\ell-1} k^{-\alpha} \\ &\leq \sum_{\ell=1}^{m} \left(\frac{et}{\ell} \right)^{\ell} 2^{\alpha(\ell-1)} \zeta(\alpha)^{\ell-1} \sum_{k=m}^{\infty} k^{-\alpha}. \qquad (4) \end{aligned}$$

The two summations are independent, so we calculate them individually:

$$\sum_{k=m}^{\infty} k^{-\alpha} \leq \int_{m-1}^{\infty} x^{-\alpha} dx = \frac{(m-1)^{-\alpha+1}}{\alpha-1}. \tag{5}$$

In addition, by lemma 3.8, we have

$$\left(\frac{e}{\ell}\right)^{\ell} \leq \frac{e}{\sqrt{\ell}(\ell-1)!} \leq \frac{e}{(\ell-1)!} \text{ for } \ell \geq 1.$$

Thus,

$$\sum_{\ell=1}^{m} \left(\frac{et}{\ell}\right)^{\ell} 2^{\alpha(\ell-1)} \zeta(\alpha)^{\ell-1} \leq \sum_{\ell=1}^{m} (t2^{\alpha}\zeta(\alpha))^{\ell-1} \cdot \frac{et}{(\ell-1)!} \leq te^{t2^{\alpha}\zeta(\alpha)+1}, \tag{6}$$

where in the second inequality we used the fact that $\sum_{\ell=1}^{\infty} \frac{x^{\ell-1}}{(\ell-1)!} = e^x$ for any $x > 0$. Inserting inequalities (5) and (6) into (4), we get

$$\mathbb{P}(T_m \leq t) \leq \frac{et}{\alpha-1} \cdot \frac{e^{t2^{\alpha}\zeta(\alpha)}}{(m-1)^{\alpha-1}}.$$

Given any $\varepsilon \in (0,1)$, we solve for

$$\frac{et}{\alpha-1} \cdot \frac{e^{t2^{\alpha}\zeta(\alpha)}}{(m-1)^{\alpha-1}} \leq \varepsilon$$

to get

$$\begin{aligned} te^{t2^{\alpha}\zeta(\alpha)} &\leq \varepsilon(m-1)^{\alpha-1}(\alpha-1)e^{-1} \\ \text{or } t &\leq \frac{(\alpha-1)\ln(m-1) - \ln(et/(\alpha-1)\varepsilon)}{2^{\alpha}\zeta(\alpha)}. \end{aligned}$$

Define $c := 2^{\alpha}\zeta(\alpha)$ and

$$t_m := \frac{\alpha-1}{c} \ln(m-1) - \frac{1}{c} \ln((c\varepsilon)^{-1} e \ln(m-1)).$$

Then we have $t_m e^{t_m 2^\alpha \zeta(\alpha)} \leq \varepsilon(m-1)^{\alpha-1}(\alpha-1)e^{-1}$ and

$$\mathbb{P}(T_m \geq t_m) \geqslant 1-\varepsilon \text{ or } \liminf \frac{T_m}{\log(m-1)} \geq \frac{1}{c} \text{ with high probability.}$$

This completes the proof. ■

We now examine the upper bound for the time T_m.

Proof of Theorem 3.2. Redefine k as a layer number such that $0 \leq k \ll m$. We first prove that all 2^k nodes in layer k are infected by time ck for some constant $c > 0$. Let $\sigma_1, \sigma_2, \ldots, \sigma_{2^k}$ be all the nodes in layer k. We have

$$\begin{aligned}\mathbb{P}(\max_{1\leq i\leq 2^k} T(\emptyset, \sigma_i) \geq t) &\leq \sum_{i=1}^{2^k} \mathbb{P}(T(\emptyset, \sigma_i) \geq t) \\ &= 2^k\, \mathbb{P}(T(\emptyset, \sigma_i) \geq t) \leq 2^k\, \mathbb{P}(X_1 + X_2 + \cdots + X_k \geq t),\end{aligned}$$

where all X_i's are i.i.d. exponential random variables with rate $r(1) = 1$. The first inequality is derived from lemma 3.11. The second inequality comes from the fact that if $\mathbb{P}(T(\emptyset, \sigma_i) \geq t)$, then $X_1 + X_2 + \cdots + X_k \geq t$ is true, but not the converse. By lemma 3.10, we have that

$$\mathbb{P}(\max_{1\leq i\leq 2^k} T(\emptyset, v_i) \geq t) \leq e^{-t}\Big(\frac{2et}{k}\Big)^k.$$

For $t = ck$, the right-hand side equals $(2ce^{1-c})^k$. We choose $2ce^{1-c} < 1$, or $c > 2.67835$, so that

$$\mathbb{P}(\max_{1\leq i\leq 2^k} T(\emptyset, v_i) \geq ck) \leq (2ce^{1-c})^k \to 0 \text{ as } k \to \infty.$$

Notice that the condition $t > k$ for lemma 3.10 is followed because $t = ck$ and $ck > k$.

Now, we perform the following procedure: by iterating k from layer 1 to layer m, and finding an upper bound on the amount of time it takes for the infection to travel from layer k to layer m using one jump, we find the minimum total time over all possible values of k and use this as the upper bound for T_m.

On average, the infection route that takes the longest amount of time from one node to a given node is the route that goes directly to the node. We know this because the mean of an exponential distribution is the reciprocal of the rate function. Note that for any $x, y \in \mathbb{Z}$ and $x \leq y$, given any of the 2^k paths from a node on layer k to a node on layer n,

$$\begin{aligned} & x^\alpha + 2^{y-x}y^\alpha \leq 2^y(x+y)^\alpha && (7) \\ \implies\ & 2^x x^\alpha + 2^y y^\alpha \leq 2^{x+y}(x+y)^\alpha \\ \implies\ & \frac{1}{r(x)} + \frac{1}{r(y)} \leq \frac{1}{r(x+y)}. && (8) \end{aligned}$$

Inequality (7) follows from adding the inequalities $x^\alpha + y^\alpha \leq (x+y)^\alpha$ and $(2^{y-x}-1)y^\alpha \leq (2^y - 1)(x+y)^\alpha$.

Now, inequality (8) implies that the mean time that it takes for the infection to travel from a starting node to an intermediate node to a final node is less than the mean time for the infection to travel directly from a starting node to a final node. In other words, given a starting node at layer n and a final node below it at layer $n+x+y$, the infection takes more time, on average, to infect the final node directly than to infect an intermediate node at layer $n+x$ and then the final node. We will thus find an upper bound on this value.

We define

$$L_{k,m} := \min_{|\sigma_0|=k, |\sigma_1|\geq m, \sigma_1 < \sigma_0} \omega_{\sigma_0,\sigma_1}.$$

We use the minimum property established in lemma 2.1 to show that $L_{k,m}$ is an exponential distribution with rate r_{total}. Because any node on layer k can directly infect 2^{n-k} different nodes at layer n through a one-jump procedure, and because there are 2^k nodes on layer k, we have

$$\begin{aligned} r_{\text{total}} = 2^k \sum_{n=m}^{\infty} 2^{n-k} r(n-k) &= 2^k \sum_{n=m}^{\infty} (n-k)^{-\alpha} \\ &\leq 2^k \int_{m-k-1}^{\infty} \ell^{-\alpha} d\ell = \frac{2^k(m-k-1)^{-\alpha+1}}{\alpha - 1}. \end{aligned}$$

The mean of $L_{k,m}$ is the reciprocal of r_{total}, or $(\alpha-1)2^{-k}(m-k-1)^{\alpha-1}$. Thus, the total infection time $\max_{1\le i\le 2^k} T(\emptyset, v_i) + L_{k,m}$ is of the order

$$t(k) = ck + (\alpha-1)2^{-k}(m-k-1)^{\alpha-1} \le ck + (\alpha-1)2^{-k}m^{\alpha-1},$$

because $k \ll m$. Then to minimize infection time to get a strong upper bound, we take

$$t'(k) = c + (\alpha-1)2^{-k}(-\ln 2)m^{\alpha-1} = 0 \text{ or } k = \log_2((\alpha-1)\ln 2\cdot m^{\alpha-1}) - \log_2 c,$$

so that

$$t(k) = c\log_2 \frac{e(\alpha-1)\ln 2\cdot m^{\alpha-1}}{c} = c\log_2\frac{e(\alpha-1)\ln 2}{c} + \frac{c(\alpha-1)}{\ln 2}\ln m.$$

Finally, we use the result that

$$\begin{aligned}\mathbb{P}(\max_{1\le i\le 2^k} T(\emptyset, v_i) + L_{k,m} \ge ck + y) &\le \mathbb{P}(\max_{1\le i\le 2^k} T(\emptyset, v_i) \ge ck) + \mathbb{P}(L_{k,m}\ge y)\\ &\le (2ce^{1-c})^k + e^{-yr_{\text{total}}}\end{aligned}$$

for $y = (\alpha-1)2^{-k}m^{\alpha-1} = \frac{c}{\ln 2}$ to complete the proof. ■

4 Illustrations

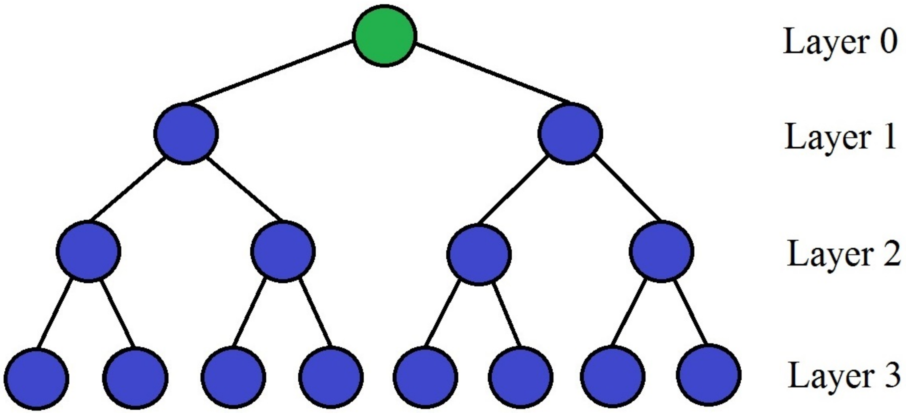

Figure 1: A binary tree with layers labeled and root node infected

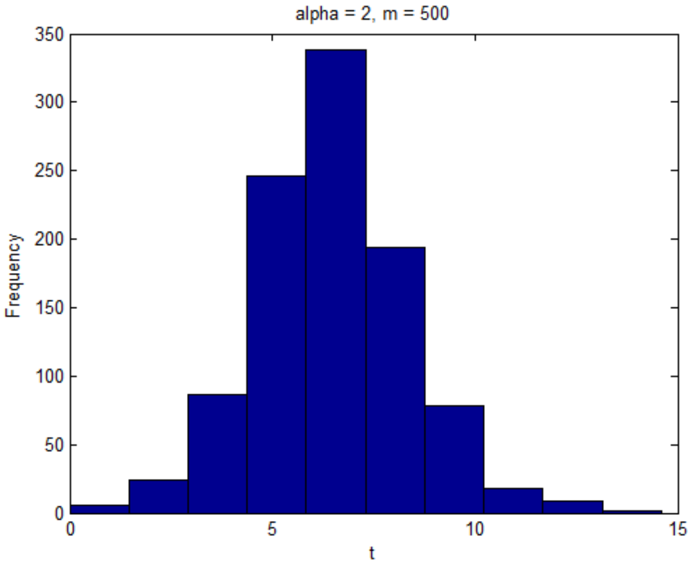

Figure 2: Simulation with $\alpha = 2, m = 500$, and 1000 trials

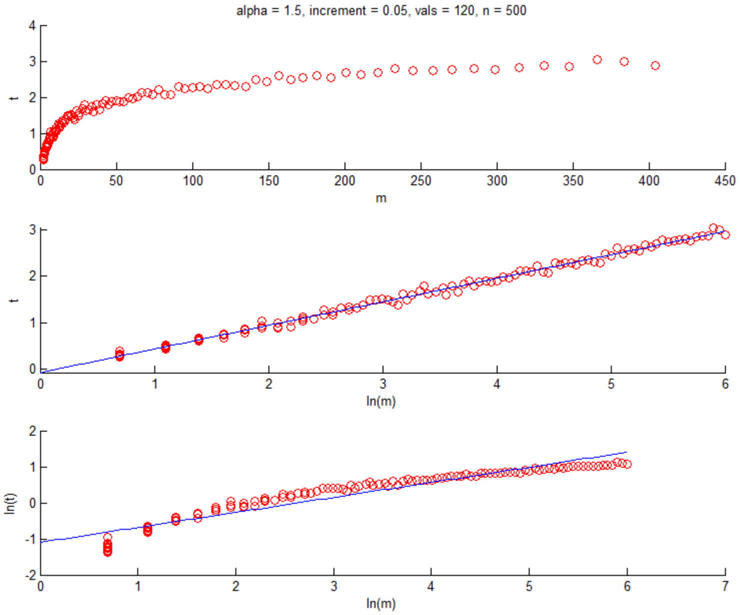

Figure 3: Regression graph for $\alpha = 1.5$, m taking on all values in $\lceil e^{0.05x} \rceil$ for $x \in \mathbb{Z}$ and $1 \leq x \leq 120$, and 500 trials per value of x.

Figure 1 is a conventional, un-directed binary tree. Figures 2 and 3 are produced in MATLAB for the purpose of simulating the spreading algorithm. Figure 2 plots the number of occurences of time t in a histogram. The shape of the outline of the histogram is approximately normal. Figure 3 plots the values of m vs. t, $\ln(m)$ vs. t, and $\ln(m)$ vs. $\ln(t)$. The line for $\ln(m)$ vs. t is more linear than $\ln(m)$ vs. $\ln(t)$, indicating the accuracy of the model which predicts that the time is a constant multiplied by $\ln(m)$.

5 New Rates

We investigate the case where

$$r(k) = \frac{1}{b^k}$$

with $b > 2$.

Theorem 5.1. *We claim that if*

$$t_m = \frac{m}{4}\Big(\ln\frac{b}{2}\Big)^2 - q\ln m \cdot \ln\sqrt{\frac{b}{2}},$$

with $q > 1$, then $\mathbb{P}(T_m > t_m)$ approaches unity as $m \longrightarrow \infty$.

As a corollary, this implies that T_m grows linearly with m.

Proof. The proof is similar to that of theorem 3.2. Using the definitions and notation from this theorem, we have

$$f_\ell(n) := \sum_{\sigma_1<\sigma_2<\cdots<\sigma_\ell:|\sigma_\ell|=n} \prod_{i=1}^{\ell} r(|\sigma_i| - |\sigma_{i-1}|).$$

Notice that the product evaluates to

$$\prod_{i=1}^{\ell} r(|\sigma_i| - |\sigma_{i-1}|) = \prod_{i=1}^{\ell} \frac{1}{b^{|\sigma_i|-|\sigma_{i-1}|}} = \frac{1}{b^{|\sigma_\ell|-|\sigma_0|}} = \frac{1}{b^n} = b^{-n}$$

Now, we can pick σ_ℓ as any of the 2^n nodes at layer n. From there, we choose $\ell-1$ out of the $n-1$ remaining nodes for a value of

$$\binom{n-1}{\ell-1}$$

As a result,

$$f_\ell(n) = 2^n b^{-n}\binom{n-1}{\ell-1}$$

We must then bound

$$\mathbb{P}(T_m \le t) \le \sum_{\ell=1}^{m}\left(\frac{et}{\ell}\right)^{\ell}\sum_{k=m}^{\infty} f_\ell(k) = \sum_{\ell=1}^{m}\left(\frac{et}{\ell}\right)^{\ell}\sum_{k=m}^{\infty} 2^k b^{-k}\binom{k-1}{\ell-1}.$$

We note that

$$\binom{k-1}{\ell-1} = \frac{(k-1)!}{(\ell-1)!(k-\ell)!} = \frac{1}{(\ell-1)!}\prod_{i=1}^{\ell-1}(k-i) \leq \frac{k^{\ell-1}}{(\ell-1)!}.$$

This implies that

$$\mathbb{P}(T_m \leq t) \leq \sum_{\ell=1}^{m}\left(\frac{et}{\ell}\right)^{\ell}\sum_{k=m}^{\infty}2^k b^{-k}\frac{k^{\ell-1}}{(\ell-1)!}.$$

Now, let $c := \frac{b}{2}$ and define $u, v > 1$ such that $uv = c$. We evaluate the second summation in the above inequality:

$$\sum_{k=m}^{\infty}c^{-k}\frac{k^{\ell-1}}{(\ell-1)!} = \sum_{k=m}^{\infty}u^{-k}v^{-k}\frac{k^{\ell-1}}{(\ell-1)!}$$

The maximum value of the expression $v^{-k}k^{\ell-1}$ is received by simply setting $\frac{d}{dk}(v^{-k}k^{\ell-1}) = 0$, or

$$\begin{aligned}-\ln(v)v^{-k}k^{\ell-1} + (\ell-1)v^{-k}k^{\ell-2} &= 0\\ -\ln(v)k + (\ell-1) &= 0\\ k &= \frac{\ell-1}{\ln v}.\end{aligned}$$

However, this only holds when $\ell > 1$. We will separate out the $\ell = 1$ case of the summation. As a result,

$$v^{-\frac{\ell-1}{\ln v}}\left(\frac{\ell-1}{\ln v}\right)^{\ell-1} = e^{1-\ell}\left(\frac{\ell-1}{\ln v}\right)^{\ell-1} = \left(\frac{\ell-1}{e\ln v}\right)^{\ell-1}$$

Plugging this back into the summation, we get

$$\sum_{k=m}^{\infty}u^{-k}v^{-k}\frac{k^{\ell-1}}{(\ell-1)!} \leq \sum_{k=m}^{\infty}u^{-k}\left(\frac{\ell-1}{e\ln v}\right)^{\ell-1}\frac{1}{(\ell-1)!}$$

for $\ell \neq 1$. Separating out $\ell = 1$ and changing the dummy variable, we have that

$$\mathbb{P}(T_m \leq t) \leq et \sum_{k=m}^{\infty} u^{-k} v^{-k} + \sum_{\ell=1}^{m-1} \Big(\frac{et}{\ell+1}\Big)^{\ell+1} \sum_{k=m}^{\infty} \frac{u^{-k} e^{-\ell}}{\ell!} \Big(\frac{\ell}{\ln v}\Big)^{\ell}$$

Making multiple simplifications, we get that this equals

$$et \frac{(uv)^{-m}}{(1-(uv)^{-1})} + et \sum_{\ell=1}^{m-1} \Big(\frac{t}{\ln v}\Big)^{\ell} \frac{\ell^{\ell}}{(\ell+1)^{\ell+1}} \frac{u^{-m}}{(1-u^{-1})\ell!}$$

Thus,

$$\mathbb{P}(T_m \leq t) \leq et \frac{(uv)^{-m}}{(1-(uv)^{-1})} + et \sum_{\ell=1}^{m-1} \Big(\frac{t}{\ln v}\Big)^{\ell} \frac{\ell^{\ell}}{(\ell+1)^{\ell+1}} \frac{u^{-m}}{(1-u^{-1})\ell!}$$

For the second term, we know that $\frac{\ell^{\ell}}{(\ell+1)^{\ell+1}} < 1$ for $\ell \geq 1$. We can simplify this as follows:

$$\begin{aligned} et \sum_{\ell=1}^{m-1} \Big(\frac{t}{\ln v}\Big)^{\ell} \frac{\ell^{\ell}}{(\ell+1)^{\ell+1}} \frac{u^{-m}}{(1-u^{-1})\ell!} &\leq et \sum_{\ell=1}^{m-1} \Big(\frac{t}{\ln v}\Big)^{\ell} \frac{u^{-m}}{(1-u^{-1})\ell!} \\ &= et \frac{u^{-m}}{(1-u^{-1})} \sum_{\ell=1}^{m-1} \Big(\frac{t}{\ln v}\Big)^{\ell} \frac{1}{\ell!} \end{aligned}$$

The Maclaurin series for e^x is simply $\sum_{i=0}^{\infty} \frac{x^i}{i!}$. As a result,

$$et \frac{u^{-m}}{(1-u^{-1})} \sum_{\ell=1}^{m-1} \Big(\frac{t}{\ln v}\Big)^{\ell} \frac{1}{\ell!} \leq et \frac{u^{-m}}{(1-u^{-1})} (e^{t/\ln v} - 1)$$

We now have that

$$\mathbb{P}(T_m \leq t) \leq et \frac{(uv)^{-m}}{(1-(uv)^{-1})} + et \frac{u^{-m}}{(1-u^{-1})} (e^{t/\ln v})$$

Now,

$$u^{-m} e^{t/\ln v} = e^{t/\ln v - m \ln u}.$$

Thus, for $t = \ln v(m \ln u - a_m)$, this evaluates to e^{-a_m}. If $a_m \geq k \ln m$ where k is a constant such that $k > 1$, then we know that the second term in the above inequality approaches 0 as $m \longrightarrow \infty$.

Now, we must maximize t in terms of u and v in order to get the best lower bound. Note that $\ln u + \ln v = \ln uv = \ln \frac{b}{2}$, which is constant. As a result, we have, by the AM-GM inequality,

$$\frac{\ln \frac{b}{2}}{2} = \frac{\ln u + \ln v}{2} \geq \sqrt{\ln u \ln v}$$

The equality case is achieved (and t is maximized) when $u = v = \sqrt{\frac{b}{2}}$, leading to $\sqrt{\ln u \ln v} = \frac{1}{2} \ln \frac{b}{2}$. Thus, we have that $t_m = \frac{m}{4}(\ln \frac{b}{2})^2 - a_m \ln \sqrt{\frac{b}{2}} = \frac{m}{4}(\ln \frac{b}{2})^2 - a_m \ln \sqrt{\frac{b}{2}}$ where $a_m \geq k \ln m$ as defined above. As a result,

$$\lim_{m \longrightarrow \infty} \mathbb{P}(T_m \leq t_m) = 0$$

which implies that $\mathbb{P}(T_m > t_m)$ approaches unity as $m \longrightarrow \infty$. ■

6 Discussion

For a binary tree with rate function $r(k) = 2^{-k}k^{-\alpha}$, we prove that when $t_m = \frac{\alpha-1}{c} \ln(m-1) - \frac{1}{c} \ln((c\varepsilon)^{-1} e \ln(m-1))$, then $\mathbb{P}(T_m \geq t_m) \geqslant 1 - \varepsilon$, meaning that the time it takes for the infection to travel from $\emptyset$ to below layer m is greater than t_m given a margin of error ε. In addition, for the upper bound, we prove that $t(k) = c \log_2 \frac{e(\alpha-1)\ln 2}{c} + \frac{c(\alpha-1)}{\ln 2} \ln m$. We have bounds for the minimum and maximum times, both on the order of $c_1 \ln(m) + c_2$ with c_1 constant and c_2 essentially constant compared to $\ln(m)$. Because we assume that m is large, these bounds limit the time that it takes for the infection to reach layer m to a very small margin of error.

We find that the time that it takes for the disease to spread below a certain layer is bounded logarithmically. Thus, growth is the inverse function of time, or exponential. In this context, growth means the maximum distance from the root node that is infected after a given time t. This fact means that near some layer number that is an exponential function of t, there is a high probability that at least one node will be infected. It is not possible to conclude the number or density of infected nodes at a specific layer. On the other hand, an estimate is obtained based on the nature of the spreading pattern. Indeed, it is very likely that one node will be infected near layer e^t. This implies that in a real life setting, after a certain time, the infection will be likely to spread to at least one person very far from the root node. The infection would also be very sparse far away from the root and more concentrated as the nodes closer to the root are examined. In other words, we consider the ratio of the number of infected people to the number of total people (2^k) at a certain layer k. This ratio grows larger as k decreases.

On the other hand, for the case where $r(k) = \frac{1}{b^k}$, we prove that when $t_m = \frac{m}{4}\Big(\ln\frac{b}{2}\Big)^2 - q\ln m \ln\sqrt{\frac{b}{2}}$, $\mathbb{P}(T_m > t_m)$ also approaches 1. As a result, the time it takes for the infection to travel from $\emptyset$ to below layer m is linear on average. As stated above, growth is the inverse function of time, or also linear.

It was proved in [3] that the bounds for the infection time are highly dependent on the values for α in the case of an infection model in $\mathbb{Z}^d$. Namely, growth is linear if $\alpha > 2d+1$, super-linear if $\alpha \in (2d, 2d+1)$, exponential if $\alpha \in (d, 2d)$, and instantaneous if $\alpha < d$. In comparison, for the model of the binary tree, the time that it takes to expand below a certain layer is logarithmic for all $\alpha > 1$. A binary tree has approximately 2^{k+1} nodes a distance of k or less away from the root node, while in $\mathbb{Z}^d$, the approximate number of lattice points a distance k or less away from the origin of the infection is on the order of k^d multiplied by a constant. This fact explains the difference between the case in $\mathbb{Z}^d$ and the binary tree. For this reason, these observations enhance the findings of other studies in the field.

7 Conclusions and Future Work

From our research, we learn for $r(k) = 2^{-k}k^{\alpha}$, that the lower bound for the time is $\frac{\alpha-1}{c}\ln(m-1) - \frac{1}{c}\ln((c\varepsilon)^{-1}e\ln(m-1))$. The upper bound for the time is $c\log_2\frac{e(\alpha-1)\ln 2}{c} + \frac{c(\alpha-1)}{\ln 2}\ln m$. As stated above, both are in the form $c_1\ln(m) + c_2$ with c_1 constant and c_2 essentially constant compared to $\ln(m)$. This confirms our hypothesis. For the case where $r(k) = \frac{1}{b^k}$, we learn that the time of infection is approximately $t_m = \frac{m}{4}\left(\ln\frac{b}{2}\right)^2 - \ln m \ln\sqrt{\frac{b}{2}}$. The infection time in this case is linear on average. In the future, it would be useful to find the approximate density and number of infections at each layer of the binary tree, along with studying perfect n-ary trees. Optimizing the algorithm to simulate the infection process is also useful, because with large α, the time that it takes the algorithm to run is quite large. Improving the algorithm is essential to verifying if the bounds are correct and replicating the results of the proof. The program will also help formulate a hypothesis in the general case of n-ary trees, and improving the running time of the algorithm will streamline this process. Questions that still remain unanswered include the time that it takes for the infection to spread to a certain layer when the binary tree is not perfect, or in the general case, an n-ary tree is not perfect. The representation of such a graph will help generalize the bounds found in this paper.

8 Acknowledgements

I would like to sincerely thank Professor Partha Dey for his guidance and support throughout this project. I would also like to extend my thanks to Dr. Tanya Khovanova for her overall guidance on the project and Dr. Slava Gerovitch and the MIT PRIMES-USA program for providing me with this research opportunity.

References

[1] Aizenman, M.; Newman, C. M. Discontinuity of the percolation density in one dimensional $1/|x-y|^2$ percolation models. *Communications in Mathematical Physics 107 (1986)*, no. 4, 611–647.

[2] Newman, C. M.; Schulman, L. S. One dimensional $1/|j-i|^s$ percolation models: The existence of a transition for $s > 2$. *Communications in Mathematical Physics 104 (1986)*, no. 4, 547–571.

[3] Shirshendu Chatterjee and Partha S. Dey. Multiple Phase Transitions in Long-Range First-Passage Percolation on Square Lattices. To appear in *Communications on Pure and Applied Mathematics*.

[4] William Feller. *An Introduction to Probability Theory and Its Applications. Vol. I.* John Wiley & Sons, Inc., New York-London-Sydney, Third edition, 1968.

[5] Sheldon Ross. *A First Course in Probability*. Macmillan Co., New York; Collier Macmillan Ltd., London, Second edition, 1984.

[6] Lyons, R.; Peres, Y. *Probability on Trees and Networks.* 2005.