PREDICTING CASCADE SIZE DISTRIBUTION ON ONE-DIMENSIONAL GEOGRAPHIC NETWORKS

Yosef Treitman

Submitted in Partial Fulfillment of the Requirements
for the Degree of

DOCTOR OF PHILOSOPHY

Approved by:
Peter R. Kramer, Thesis Adviser
Mark Holmes, Member
Ronie Lai, Member
Gyorgy Korniss, Member

Department of Mathematics
Rensselaer Polytechnic Institute
Troy, New York

[August 2018]
Submitted July 2018

CONTENTS

LIST OF FIGURES

NOTATION USED IN THIS WORK

- $\mathbf{A}$ is the adjacency matrix of the network
- a is the time interval that we analyze when using the regression-based approximation.
- a_u is the likelihood that an active agent will immediately cause exactly u other agents to activate in a branching process.
- $\mathrm{Bin}(x, p, N)$ is the probability mass function at x of a binomial random variable with N trials and success probability p.
- C is the clustering coefficient, or the probability that the neighbor of a neighbor of an agent is also a neighbor of that agent.
- $d(n_i, n_j)$ denotes the spatial distance between agents n_i and n_j. Not to be confused with the shortest path length between two nodes, often referred to as "distance" in the study of non-spatial networks.
- F is the cumulative distribution function of the response thresholds of the agents in the network.
- G is the cumulative distribution function of the final cascade size.
- G_{spon} is the cumulative distribution function of the final cascade size under the assumption that only spontaneous terminations are possible.
- G_{exst} is the cumulative distribution function of the final cascade size under the assumption that spontaneous terminations are impossible and the only source of termination is exhaustion of the supply of available agents.
- H is the cumulative distribution function of the final cascade time.
- k is the number of neighbors of an agent.
- L is the cloning factor of a cloned network.
- l is the mean intervertex path length between all pairs of agents in a network.

- l_i represents the level of an agent in a tree. If an agent is at level l_i, that means the shortest path length between that node and the root of the tree has length i.
- m denotes the number of already active neighbors of a given node.
- M_0 is the size of the microcosm we use when estimating the behavior for the whole network.
- N is the size of the network and is equal to the number of agents in the network. Not to be confused with w, the width of the geographic map in which the network is embedded.
- N_0 is the number of initial seeds in a network.
- $n_1, n_2, n_3...$ are the individual agents in the network.
- n is the number of recently activated neighbors of an agent.
- $\mathbf{P}$ is the probability transition matrix of the number of agents that activate at a given time.
- P_{Conn} is the probability that a given pair of recently activated agents are adjacent.
- $P_e(n_i, n_j)$ is the probability that agents n_i and n_j are connected by an edge.
- P_{range} is the probability that a given pair of recently activated agents are separated by less than one radius of influence.
- p is the probability that any pair of agents whose distance is less than the radius of influence r are adjacent.
- p_k is the fraction of nodes in the network that have degree k.
- $p(n|k, m, x, \tau - 1)$ is the probability that an inactive agent at location x received n new spikes at time $\tau - 1$ given that it had k neighbors and m already active neighbors.
- $p(n|k, m, x, \tau - 1, y_{\tau-1})$ is the probability that an inactive agent at location x received n new spikes at time $\tau - 1$ given that it had k neighbors and m already active neighbors and that $y_{\tau-1}$ total agents activated at time $\tau - 1$.

- $\mathbf{q}$ is the quasistationary distribution of Y_τ.
- q represents the likelihood that a neighbor of a node activates, given that the node itself is still inactive.
- q_n is the probability that an agent in a tree at level $d-n$ will be active if its parent is inactive, where d is the highest level of the tree.
- $\bar{q}(x,\tau)$ represents the likelihood that a neighbor of a node at point x was active at time τ, conditioned on its being inactive at time $\tau-1$.
- $R(x)$ is the region which is in range of a given agent located at x.
- r is the radius of influence of each agent in the network. If the spatial distance between two nodes exceeds r, then those two nodes cannot be adjacent.
- $T_\tau(k,m,t,x)$ is the probability that an agent that was inactive at time $\tau-1$ at point x has degree k, had m active neighbors at time $\tau-1$, and had threshold t.
- $T_{\tau,y_{\tau-1}}(k,m,t,x)$ is the probability that an agent that was inactive at time $\tau-1$ at point x has degree k, had m active neighbors at time $\tau-1$, and had threshold t, given that $y_{\tau-1}$ total agents activated at time $\tau-1$.
- $\hat{T}_\tau(k,m,t,i)$ is the average value of T at the appropriate values of k, m, t, and $i-1 < x \leq i$ at time τ.
- t is the response threshold of a given agent.
- t_{mod} is the average response threshold of agents that activate.
- $U(y_{\tau-1})$ is the average number of spikes sent from each agent that activated at time $\tau-1$ to agents that were still inactive at time $\tau-1$ given that $y_{\tau-1}$ total agents activated at time $\tau-1$.
- $U_{\text{emp}}(y_{\tau-1})$ is the empirically determined average value of $U(y_\tau-1)$.
- $U_{\text{pred}}(y_{\tau-1})$ is the predicted value of $U(y_\tau-1)$ based on a single time step of our simulation.

- $U_{\text{theor}}(y_{\tau-1})$ is the theoretical value of $U(y_{\tau-1})$ based on analysis of the number of neighbors of an agent that activated at time $\tau - 1$ that are likely to have activated before time $\tau - 1$, at time $\tau - 1$ exactly, or still be inactive at time $\tau - 1$.

- u_i is a random variable following a uniform distribution between 0 and 1. u_i represents the position of agent n_i relative to the left boundary of its interval.

- $V(y_{\tau-1})$ is the average number of spikes received by each agent that activated at time $\tau - 1$ in excess of its response threshold given that $y_{\tau-1}$ total agents activated at time $\tau - 1$.

- $W(y_{\tau-1})$ is the expected number of spikes sent by each agent that activated at time $\tau - 1$ to other agents that activated simultaneously with itself over a single time step given that $y_{\tau-1}$ total agents activated at time $\tau - 1$.

- $W_{\text{pred}}(y_{\tau-1})$ is the predicted value of $W(y_{\tau-1})$ based on a single time step of our simulation.

- $W_{\text{theor}}(y_{\tau-1})$ is the theoretical value of $W(y_{\tau-1})$ based on analysis of the number of neighbors of an agent that activated at time $\tau - 1$ that are likely to have activated before time $\tau - 1$, at time $\tau - 1$ exactly, or still be inactive at time $\tau - 1$.

- w is the spatial width of the network. It is a measure of the size of the map on which the network is embedded.

- x is the geographic coordinate of an agent.

- Y_τ is the number of agents that activated at time τ.

- z is the final cascade size.

- μ is the nominal mean degree of the agents in a network.

- $\mu_{\text{eff}}(x)$ is the expected degree of an agent at a specified point in the network. Usually, $\mu_{\text{eff}}(x) = \mu$, but for nodes sufficiently close to the boundary of the map, $\mu_{\text{eff}}(x) < \mu$.

- μ_{mod} is the mean degree of agents that activate.

- ρ is the final cascade size as a fraction of the total size of the network. Contrast with z, which is the final cascade size as a number of agents.
- ρ_{sat} is the final cascade fraction of a saturated cascade.
- $\rho(x, \tau)$ is the probability that an agent at point x activated at or before time τ.
- $\rho_{y_\tau}(x, \tau)$ is the probability that an agent at point x was active at time τ, given that y_τ agents activated at time τ exactly.
- $\rho_{\text{new}}(x, \tau)$ is the probability that an agent at point x activated at time τ exactly.
- $\rho_{\text{new},y_\tau}(x, \tau)$ is the probability that an agent at point x activated at time τ exactly, given that y_τ agents activated at time τ exactly.
- ρ_0 is the initial seed size as a fraction of the total network size.
- $\sigma(y_{\tau-1})$ is the sample standard deviation of activation locations given that $y_{\tau-1}$ agents activated at time $\tau - 1$ exactly.
- τ is the time that has elapsed during the cascade, or the number of time steps since the cascade began.
- τ^* is the time at which we assume that the wavefront stabilized.
- $\omega(x, \tau - 1)$ is the relative activation weight of the neighbors of an agent at location x that were not already active at time $\tau - 2$, compared to the non-neighbors of that agent.

ACKNOWLEDGMENTS

I must express my gratitude to Dr. Peter R. Kramer for agreeing to oversee this project. I greatly appreciate how Professor Kramer has balanced my independence with my need for guidance.

Additionally, I owe thanks to Professors Mark Holmes, Ronjie Lai, and Gyorgy Korniss for taking the time to serve on my thesis committee.

Finally, I owe thanks to the National Science Foundation for five semesters and three summers of funding. This work was partially supported by NSF RTG grant DMS-1344962

ABSTRACT

Over the past few decades, there has been considerable research on the spread of various phenomena across networks. While the most general case of the cascade problem on an arbitrary network is too broad a question to address, the question has been studied under specific simplifying assumptions, both on the construction of the network and on the rules for the spread and adoption of the cascading phenomenon. In paricular, we are interested in studying cascades under the widely-used Centola-Macy threshold model on a class of spatial networks [3]. The spatial aspects of these networks present challenges not found in other networks.

We make rather specific assumptions about the distribution of agents across the geographic map and the likelihood of any pair of them being adjacent. We assume that the agents are nearly evenly spaced across the network. More precisely, we assume that if a network of width w were divided into N equal intervals of width $\frac{w}{N}$ each interval would contain exactly one agent. The location of each agent within its interval follows a uniform distribution. We assume that each agent can only be adjacent to other agents within some radius of influence r of itself with $r \ll w$. As the cascade propagates across the spatial network, it may spontaneously terminate. This property is not found in locally treelike networks [12, 27]. This work examines the likelihood of such a termination and accounts for that possibility when estimating the CDF of the final cascade size distribution.

To address the main challenge of the possibility of a finite-time extinction, we find the mean number of new activations per unit time and the likelihood of a spontaneous termination. These statistics can be estimated by viewing the number of new activations per unit time as a Markov chain. Given the number of activations at some time $\tau - 1$, we can estimate the number of spikes sent to inactive agents. Given that number, we can estimate the number of activations at time τ. We assume that the number of new activations follows a Poisson distribution, and use our estimate as the mean. This gives us an approximate one-step probability transition matrix of the number of new activations from time $\tau - 1$ to time τ. Using this matrix, we find the mean number of new activations and the extinction probability.

The approximation we develop is reasonably accurate for several response threshold

distributions. Previous work already showed that, under these assumptions, the cascade will propagate at a constant overall speed [4, 36, 38]. Our approach gets an estimate for that speed and the likelihood of spontaneous termination.

CHAPTER 1
INTRODUCTION

The phenomenon of changes cascading across a network has been studied heavily. Because the precise rules governing how the cascade will propagate are unknown, some simplifying assumptions need to be made. A commonly used set of assumptions is the Centola-Macy model [3]. Under this model, an agent on the network will activate and become part of the cascade once a large enough number of its neighbors have activated. The necessary number of active neighbors is called the *response threshold* of the agent. Once active, an agent can never deactivate. Recently, a great deal of research has been conducted addressing the question of when large cascades may occur or will occur [1, 3, 8, 9, 10, 11, 12, 13, 14, 15, 16, 20, 21, 23, 24, 26, 27, 28, 30, 32, 33, 34, 37, 38, 39, 40, 41]. This work seeks to build upon those findings by addressing the possibility that an active cascade will spontaneously stop. Thus, we seek not only to approximate the probability that a cascade will be large rather than small, but we estimate the cumulative distribution function of the final cascade size. We build on the work of Watts, Gleeson, Cahalane, Newman, Porter, Mollison, Daniels, and many others in the study of cascading epidemics.

There are several network statistics which have noticeable relevance. Prior work discusses the relevance of the degree distribution and the response threshold distribution of the agents in the network [2, 9, 10, 11, 12, 13, 14, 27, 31, 33, 34, 35, 38, 40]. As one might expect, if the response thresholds are lower overall, so that agents activate more easily, the expected cascade size will increase. Similarly, if the mean degree is higher, so that each active agent will have the opportunity to activate more agents, the expected cascade size will increase. Further topological concerns include the likelihood of clustering, where agents are more likely to be adjacent if they have at least one mutual neighbor, and the mean intervertex path length of the network [8, 10, 13, 14, 15, 26, 27, 28, 30, 31, 32, 38, 41]. An approach to addressing these challenges could be to ignore those complicating factors, assuming that the network behaved like an Erdos-Renyi network (where clustering is not a concern) or like a tree (where clustering is impossible). This assumption that the network is "locally tree-like" is effective on a wide class of networks [27]. This work addresses a class of networks where the assumption fails.

The class of networks considered in this model relies heavily on geography. It is assumed that the network is embedded in a one-dimensional spatial map of width w. Each agent can only be adjacent to other agents within some radius of influence r of itself. We assume that $r \ll w$. A trait reflected in cascades on such networks that is not found on locally treelike networks is the realistic possibility of the cascade terminating after it has grown to some arbitrary size. On locally treelike networks, either the cascade terminates almost immediately or it envelopes the majority of the network [12, 27]. On this class of spatial networks, the possibility of extinction must be considered at each time step. This work seeks to address that consideration.

We present two approaches to addressing this issue. In our first method, we assume that extinction probability will go through three phases. In the first brief phase, the cascade propagation is heavily dependent on the size of the initial seed. If there were a large number of initially active agents, the cascade is likely to continue, and if there were only a small number, the cascade is more likely to terminate. There comes a point where the perpetuation of the cascade depends less on the initial condition and more on the susceptibility of the network to cascade. Finally, there comes a point where the cascade is forced to terminate due to exhausting the supply of available agents. This would mean that the spontaneous extinction hazard function should be constant over time, with the exception of a boundary layer at the beginning and a boundary layer at the end. Consider the case where the cascade propagates to the point of exhausting the available agents. Some agents will have response thresholds larger than their degrees, so they will never activate. In this case, we can approximate the fraction of agents which resist this cascade by modeling the network as a tree with the same degree distribution. Other aspects of the geography have little relevance to which agents would resist this nearly-complete cascade. To address the boundary layer at initiation, we take advantage of the geographic nature of the network and restrict or attention to the subnetwork that can be affected by the initial seeds over this brief window. Because this subnetwork is so small, we can sequentially generate many such subnetworks in our class of networks and gather statistics on how often the cascade terminates after each of the first few time steps as well as the average cascade size conditioned on the termination times. The larger challenge is to estimate the likelihood of spontaneous extinction outside of those boundary layers. We do this by taking a slightly larger subnetwork that is still small enough to allow for numerous direct simulations. We can gather the same statistics of

extinction probability and conditional cascade size, and use those statistics to estimate the CDF outside the two boundary layers. Finally, we can form the composite CDF by using the fact that for the cascade to continue past a given point, it cannot have terminated for any of the possible reasons. This approach has the advantage of accuracy but requires numerous simulations to make a single prediction.

The second approach seeks to circumvent the need for a large number of numerical simulations. We can use the results of the locally treelike approximation to estimate the likelihood of a fast extinction, the conditional size of such an extinction, and the distribution of sizes of exhaustion-terminated cascades. The main challenge is determining the cascade propagation speed and the probability of a spontaneous termination during a given time step. Our approach views the number of new activations during each time step as a Markov process. It is assumed that the number of activations in a given time step can depend on the number of activations in the immediately preceding time step, but not on any history prior to that. We build the probability transition matrix associated with that Markov chain and use that matrix to find the likelihood of termination and the average number of new activations conditioned on no extinction. These statistics can be used to infer the behavior of the cascade size outside the boundary layers, and can be combined with the early-cascade and nearly-full-cascade behavior to construct the composite CDF of final cascade size. This distribution of cascade sizes reflects the numerical simulations well.

A complicating factor that needs to be considered is the number of agents that a given agent can realistically influence to activate. Suppose an agent has k neighbors. When that agent activates, it can only influence those k agents to activate. However, it cannot influence all k of them because it is possible that some of them would already be active. If the agent had response threshold t, then it would have at least t active neighbors by the time it activated. Because we use a discrete-time model, it is possible that the number of active neighbors will overshoot the response threshold, increasing the number of neighbors that were already active by the time the agent activated. Additionally, it is possible that some of its neighbors would have activated simultaneously with itself. These neighbors could not have influenced the agent itself to activate. These agents would not be counted toward the number of active neighbors until the agent was already active. We implement these considerations into a single algorithm and compare the resulting prediction to the simulations on a class of reasonably small networks.

CHAPTER 2
FUNDAMENTAL CONCEPTS REQUIRED TO MOTIVATE THE PROBLEM STATEMENT

In recent years, the field of network science has been increasingly used to model real-world phenomena [31, 35, 42]. In some select cases, there are natural models and equations which can be rigorously justified to describe the phenomena. More often, some approximations need to be made to develop useful equations. This is especially true when the processes in question involve human decision making, as this can be difficult to predict rigorously. Section 2.1 discusses some examples of these phenomena and how to fit equations to the resulting networks.

Among the questions one can ask about a network is "How likely is it that a phenomenon will spread to a given extent across the network?" This spreading is known as *cascading*. This work restricts itself to those processes which can be approximated by threshold models for cascading epidemics. We assume that a member (or *node* or *agent*) of the network will adopt the phenomenon and become part of the cascade if enough of its neighbors are already part of the cascade. These threshold models are described in more detail in Section 2.2.

Naturally, the extent to which the cascade will spread will depend on the structure of the network [8, 10, 13, 14, 15, 26, 27, 28, 30, 31, 32, 38, 41]. It is unreasonable to expect the analyst studying the network to know the all the details of the network. Furthermore, any algorithm that requires and uses such specific details would probably be intractable to implement for large networks. Therefore, we restrict our attention to a few network statistics. For example, while we may not know exactly which pairs of agents are adjacent, we do presume knowledge of the average number of neighbors of each agent and the probability distribution of the number of neighbors of the agents on the network. While we do not presume knowledge of every agent's internal activation threshold, we do presume knowledge of the probability distribution of these thresholds. These assumptions are further described in Section 2.3. Previous work is summarized in Chapter 3 and our new contributions are discussed in the later chapters.

2.1 The Relevance of Network Models and Network Representations

There are many real-world phenomena which can be modeled as processes on networks [31, 35, 42]. A linguist may want to study how one's speaking mannerisms are affected by the diction of his or her peers. A marine biologist may wish to study how an entire school of fish can swim together without separating from each other. A neuroscientist may study the effects of individual neurons on each other. An entrepreneur would be interested in the likelihood of a new product gaining widespread use. As a heuristic rule, a process can be modeled through use of a network if the following properties hold:

1) The process involves the behavior of numerous distinct items

2) The behaviors of the items can affect each other

3) Each item can only influence a selection of the other items

In each of the examples mentioned above, while the scientist may be interested in a single-parameter average, such as the total number of people who use the new product or the mean group velocity of the school of fish, the dynamics include variables relating to each individual member of the group. (The members of the network are called *agents*.) For example, each fish has its own position and velocity at each point in time. Each prospective customer of the entrepreneur has bought and maintains some quantity of the product. There is also an assumption that some agents can affect other agents in some way. In the linguistic example, suppose we are interested in whether a given person refers to a sweetened carbonated beverage as "pop," "soda," "coke," or by some other name. It is often assumed that people are more likely to use the name that they have heard used more often. In the neuroscience example, there are assumptions about how a firing neuron affects another neuron on the other side of a synapse. Finally, there is the matter of each agent affecting a selection of the others. In the marine example, it can be assumed that each fish is more heavily influenced by the behaviors of nearby fish (to avoid collisions) than fish further away. In the linguistic example, people are more likely to be influenced by people with whom they have regular contact. In the neuroscience example, a neuron can only directly affect other neurons with which it shares a synapse.

Given a physical setup which obeys the three properties listed above, an analyst may try to construct some network to represent the phenomenon. Each of the agents would be represented by a node of the network. In a visual representation of a network, the nodes

are depicted by circles. Such a visual representation is called a *graph*. Unfortunately, the term *graph* can also be used to describe a visual comparison of two or more variables on a set of axes, such as a parabola being the graph of a function $y = x^2$. This work uses both types of graphs. To avoid confusion, the term *graph* is used exclusively to describe a visual representation of a network, and the term *plot* is used to describe a visual representation of the relationship between two or more variables. At each point in time, each node may or may not be able to affect each other node. To distinguish between these possibilities, we connect the nodes that are capable of influencing each other by an edge. In the graph of a network, each edge is represented by a line segment connecting the two relevant nodes. In this work, it is assumed that the ability of agents to affect each other is symmetric. That is, if agent n_1 can affect agent n_2, then agent n_2 can affect agent n_1. For this reason, in the graphs shown in this work, if there is an edge between two nodes, then the agents represented by the two nodes can influence each other. Such graphs are called *undirected* graphs. There has been some study of cascades on asymmetric graphs [17]. While our methods may be extendable to such networks, we do not consider those cases here. Two agents which can affect each other are said to be *connected* or *adjacent* to each other. The terms *connected* and *adjacent* can also be used to describe two nodes which have an edge between them. Each agent is said to be a *neighbor* of the agents adjacent to itself, and each node is said to be a neighbor of the nodes adjacent to itself. In the most general case, some pairs of adjacent agents may be more likely to affect each other than other pairs of adjacent agents. This would be reflected in the graph by assigning some number to each edge corresponding to the relative likelihood of a pair of adjacent agents affecting each other, and that number is called the *weight* of the edge. However, this work is limited to the special case where all edges have the same weight, though not all pairs of nodes have the same probability of being joined by an edge. A network of this type is called an *unweighted network*. By convention, we assign a value of 1 to each edge, and the graph omits the weights entirely, with a weight of 1 implicitly assigned to each edge.

In addition to a network having a visual representation (in the form of a graph) there is a numerical representation as well. This numerical representation takes the form of a matrix $\mathbf{A}$ of size $N \times N$, where N represents the total number of agents in the network. If agent n_1 is not adjacent to agent n_2, then $a_{n_1,n_2} = 0$, otherwise, a_{n_1,n_2} is equal to the weight of the edge connecting the nodes corresponding to n_1 and n_2. Because this work is concerned only with

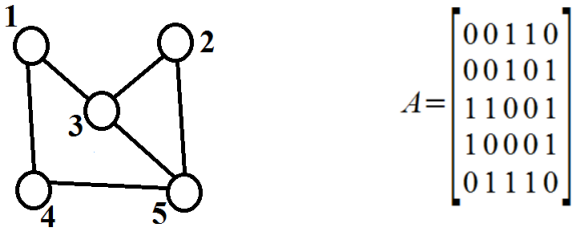

Figure 2.1: A graph representing a network (left) and its corresponding adjacency matrix (right).

unweighted networks, each entry of **A** is either 0 or 1. Additionally, because this work only concerns networks where the ability of agents to affect each other is symmetric, the matrix **A** is a symmetric matrix. The matrix **A** is referred to as the *adjacency matrix*. To illustrate the relationship between a network, that network's graph, and that network's adjacency matrix, Figure 2.1 gives an example of a small network and shows its corresponding graph and adjacency matrix. On this small network of five agents, the following pairs of agents are adjacent: Agent 1 is adjacent to agents 3 and 4, agent 2 is adjacent to agents 3 and 5, agent 3 is adjacent to agents 1, 2, and 5, agent 4 is adjacent to agents 1 and 5, and agent 5 is adjacent to agents 2, 3, and 4. The graph on the left is the graph of this network and the matrix **A**, defined on the right, is the adjacency matrix of this network. Note that **A** is symmetric, and that each entry of **A** is either a 0 or a 1.

Constructing the appropriate network for a given physical phenomenon could present its own set of challenges. Sometimes, gathering the data on exactly which agent can affect which other agents is too time consuming a task. In the neuroscience example, there could be an astronomical number of neurons, each possibly sharing a synapse with each other one. Looking at these microscopic cells to determine exactly where the synapses are is far from a trivial task. The challenge of constructing the appropriate network becomes even more difficult if there is no intuitively obvious way to determine whether or not two agents are adjacent. In the marketing example, whether one person can influence another to use the product is almost a matter of opinion. This work does not concern itself with the question of determining whether or not two nodes are adjacent, but it does not presume that the network structure is known exactly, either. Rather, this work assumes that the probability

of any two given agents being adjacent is known. In a real-world scenario, this reduces the challenge from cataloging all pairs of adjacent agents to estimating the probability that any pair of agents would be adjacent, perhaps by collecting a reasonable sample of pairs of adjacent agents.

2.2 The Centola-Macy Model for Cascades

For each physical phenomenon with an underlying network, there are a number of questions we could ask. This work focuses in particular on the question of cascading, building on the results found in [8, 9, 10, 11, 12, 13, 15, 27, 29, 34, 35, 38, 40]. As mentioned, we are interested in some parameter associated with each agent. In the marketing example, we could be interested in the quantity of product that each person maintains, or even just whether he or she maintains any product at all. In the linguistic example, we are interested in the name each person uses for a given object. (In the latter case, while there is no physically natural way to assign a number to each possible name, we can do so arbitrarily.) Different agents with the same value of the relevant parameter can be said to be in the same state. The cascade problem assumes some initial distribution of states across the agents in the network (usually with only a small number of agents in a particular state of interest) and attempts to predict how the distribution of states will change over time [9, 10, 11, 12, 13, 15, 27, 31, 34, 35, 40].

In order to predict the evolution of the cascade over time, it is not enough to know how the agents in the network are connected to each other. We also need a set of rules governing how an agent will change its state based on the states of the other agents in the network [9, 11, 12, 15, 17, 31, 34, 35, 40]. We can consider this a set of rules governing how the states of the agents will update themselves. The appropriate update rule will depend on the physical phenomenon that the analyst is trying to model with the network. For computational simplicity, we use a discrete-time update rule. This means that at fixed points in time, each agent could change its state. Note that there is a distinction between an agent updating and that agent changing its state. At a fixed series of points in time, all agents update synchronously. When an agent updates, its updated state may be the same as its previous state. An agent will only change its state if the update rule dictates that it should change is state. This work focuses on a specific class of update rules, such as those developed in [40] and [3]. The assumptions we make are the following:

1) An agent of one state will change to another state if and only if it has sufficiently

many neighbors of that state. Otherwise, it will retain its original state. Note that the condition for a node changing its state is based on the absolute number of its neighbors of the new state, as described in [3], not the proportion of its neighbors which have the new state, as described in [40]. While the latter assumption is interesting as well, it is not the focus of this work.

2) One state is considered the *active* state. Agents in the active state can never change from the active state. This condition is known as the *permanent activation property.* While this may initially seem like an arbitrary and physically inaccurate assumption, there are many cases where it would be considered a decent approximation. For example, while it is theoretically possible for a cell phone user to terminate his or her contract and never use a cell phone again, this happens rather infrequently. In the study of neuroscience, while neurons which have fired will revert back to the inactive state, if the period of interest is shorter than the refractory period of the neurons in question, neurons that have fired can be considered "already fired" and cannot revert back to the status of "capable of firing if the proper threshold is reached" during the period of observation. In cases where the time frame of observation exceeds the time scale of adoption, one may need to account for the fact that several spikes sent around the same time will be more effective than spikes that are more spaced out [1, 21]. This can lead such networks to have synchronized or "bursty" cascades [1, 21, 30]. While this work does not account for such long periods of observation, it can possibly be applied to analyzing a single burst of activity on the network.

3) For a given inactive agent of the network, the number of active neighbors it needs to activate, called the *response threshold* of the agent, does not change over time. Each agent's response threshold is a random variable following a pre-specified *response threshold distribution.* We let F represent the CDF of the response threshold distribution. In the case studied in this work, $F(0) = 0$, indicating that there are no spontaneous activations without influence from other active agents. Different agents can have different response thresholds, but a single agent will have the same response threshold at one update time as it will have at any other update time. When an agent activates, it brings all of its inactive neighbors closer to their thresholds. When this happens, we say that the agent sent a *spike* to all of its neighbors. This terminology comes from neuroscience, where a neuron sends a sudden impulse to change the voltage of its corresponding post-synaptic neurons.

4) Once an agent's response threshold is known, the state of an agent at some time τ

can be determined exactly from its own state one time step prior at $\tau - 1$ and the states of its neighbors at time $\tau - 1$. Other agents that are not adjacent to a given agent have no direct affect on the state of that given node. Additionally, the new state is not considered a random variable if the state of the network before the update is determined. More general threshold models allow for the probability of a known agent with known adoption threshold to increase steadily an increasing number of active neighbors [39]. Instead, we focus on the special case where the probability of adoption is a unit step function of the number of active neighbors.

5) All agents are presumed to update simultaneously, not one at a time. This is relevant because if a single agent updates, this may affect whether its neighbors have sufficiently many neighbors of a given state to change state. This possible effect is ignored until after all nodes have updated. Then, after all the agents have updated, the new states of the agents will be used when the next update time occurs.

This set of rules is collectively referred to as the *Centola-Macy model*, and is described in [3]. Other factors that are often considered in the study of social networks are broadcasting, where a source can transmit information to all agents, and the possibility for an agent to act on the behavior of the collective in addition to the behaviors of the individual neighbors [16, 42]. We do not consider these factors here.

In our work, we assume there to be only one active state. There are not competing active states, such as competing social ideologies or music tastes. If there are similar options for an agent to choose, we are only concerned with the agent choosing to adopt any option. For example, we would be interested in the agent's decision on joining social media, but not on the decision between Facebook and MySpace. To adapt this model to a case with multiple cascades that can overwrite each other [28], we would restrict our attention to a time frame small enough that only one cascade would be relevant at a time.

At this point, it would be beneficial to distinguish between the study of cascading epidemics (which we have been discussing until now) and cascading failures (as described in [7, 22]). In a cascading failure, an active agent will send a spike of some magnitude divided between its neighbors, rather than a single spike to each neighbor. Examples where a cascading failures model is more accurate include the study of overloading power grids and traffic backing up across highways. These cases involve the flow of some object (current, fluid, traffic, etc.) which allows for tools such as Kirchoff's law for current flow. In such a

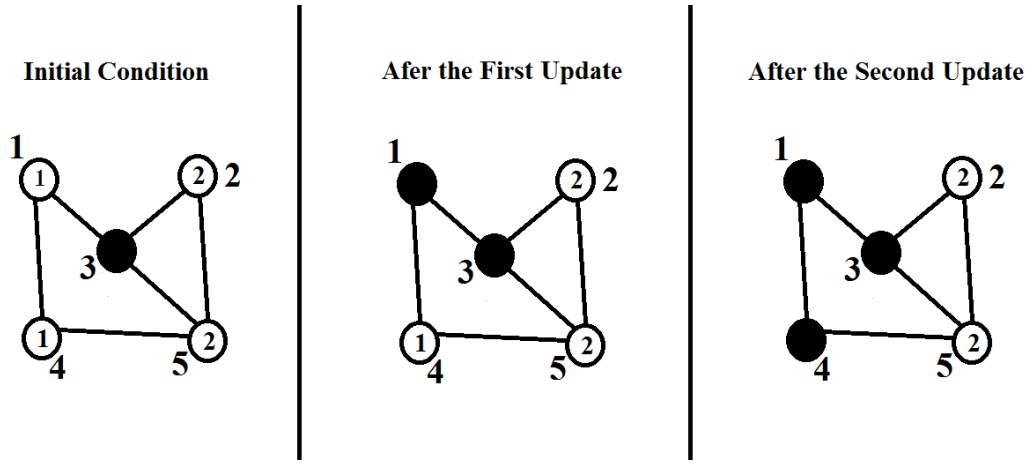

Figure 2.2: Progression of a cascade using the Centola-Macy model on the small network from Figure 2.1.

model, adding an edge can either increase the cascade size by adding an additional spike, or decrease it by decreasing the sizes of the other spikes. There has been some effort to relate the two or study the two phenomena together [19]. Our model focuses solely on cascading epidemics, and leaves the spike-size rescaling applicable to cascading failures for future work.

Figure 2.2 illustrates these rules, displaying the state of a network in its initial state, after the first update, and after the second update. In this visual representation, the numbers outside the circles represent the identification numbers of the nodes, while the numbers inside the circles represent the response thresholds of the nodes. Filled-in black nodes are active and the other nodes are inactive. While the response thresholds of active nodes are obscured by the black filling, this is not a problem because a result of the permanent activation property is that the response thresholds of active nodes have no effect on the evolution of the network state over time. The left panel represents the initial condition. The center panel represents the state of the network after the first update. Because node 1 only needs one active neighbor to activate, it becomes active after the first update. Note that node 4 does not activate at this point. If the update rule called for the update of nodes one at a time in the order of their identification numbers, node 1 would have activated before we checked to see how node 4 would update. However, because we use a simultaneous update rule, we check for whether node 4 activates while node 1 is still inactive. The right panel represents the state of the network after two updates. Because node 1 was already active immediately before the second update, node 4 has an active neighbor, and activates during the second

update.

Under the rules of the Centola-Macy model, the specific response thresholds of the individual agents of the network affect the evolution of the cascade. Determining these response thresholds can be tedious, and may not even be feasible to measure. In the marketing example, the psychology behind what will convince a person to buy a product is an inexact science. In the neuroscience example, the response threshold of a neuron depends on its initial voltage as well as some measure of how much the voltage will change when one of its neighbors fires. For this reason, it is impractical to require knowledge of the exact response threshold of each agent of the network to predict the evolution of the cascade. Instead, this work presumes knowledge of the probability of an agent having any given response threshold. These probabilities could be estimated through random sampling or through some theoretical belief about the most likely distribution of response thresholds. It is also assumed in this work that the probability distributions of the response thresholds of the different agents are equal across the different agents, and are independent of each other and of the other properties of the agents. This includes independence of the total number of neighbors of each agent, making this model different from one where it is the fraction of active neighbors, described in [40]. This has the unintuitive result that inactive neighbors of an agent do not encourage that agent to remain inactive. Thus, agents with more neighbors are neither more nor less likely to activate upon having any given number of active neighbors. There is another model where it is the fraction of one's neighbors that determines whether or not a node activates, not the absolute number of active neighbors. This model, known as the *Watts threshold model*, has been studied in [9, 11, 12, 20, 34, 37, 40].

2.3 Assumptions on Network Structure

While the exact structure of the network is difficult to determine precisely, the network structure affects the evolution of the cascade strongly enough that some properties of the network structure need to be accounted for. A rather simple model is to assume that each pair of agents has the same probability of being connected as each other pair of agents and that the connections between agents are independent of each other. A network constructed from such a model is called an *Erdos-Renyi network*, and its graph is called an *Erdos-Renyi graph* [31, 35]. The cascade question has been studied extensively on graphs of this type, and previous results are discussed in the next chapter. Realistically, there is reason to believe

that some agents would be more likely to have influence on some specific agents over others. In particular, one may suspect that a given agent would be more influenced by other agents which share some particular properties with itself. In the linguistic example, it is natural to assume that people are more likely to be influenced by the speech patterns of other people who live nearby, introducing a spatial component to the network [38]. In the neuroscience example, while we may not know exactly where the neuron synapses are, we may assume that there are more connections between neurons in the same part of the body than in distant parts of the body. We can consider all relevant details of the nodes to be mappable to some higher-order metric space. For example, we may presume that people are socially more likely to influence other people of similar location, age, and gender. Under that assumption, we can consider each agent to have a specific set of coordinates in the four-dimensional metric space of latitude, longitude, age, and gender, and generate some rule about the likelihood of two agents being neighbors if they are within some distance of each other in the four-dimensional metric space. In general, such a higher-order metric space can be considered a higher-order geography. It has been shown that on networks where the connections heavily depend on the underlying geography, using the technique intended for Erdos-Renyi networks has limited success [27, 38]. The successes and shortcomings of this approach are discussed in the following chapter.

This work is concerned with the those networks where the approach meant for Erdos-Renyi networks is ineffective due to there being some geographic reason for some pairs of agents being more likely to be connected than others. While it would be ideal to address problems where the underlying geometry had any number of dimensions, this work only focuses on those networks with a single relevant geographic coordinate, in the hopes that it will pave the way for solving similar problems with higher-dimensional underlying geography.

Regardless of the number of geographic dimensions of the network, we need to account for how the geographic distance between two agents affects the likelihood that they will be connected. This work focuses on the simple case where agents can only be neighbors if they are sufficiently close, and any pair of sufficiently close agents is equally likely to be connected as any other. More formally, if we denote the maximum allowable distance of connected nodes by r and call it the *radius of influence*, denote the distance between two nodes n_1 and n_2 by $d(n_1, n_2)$, and the probability that nodes n_1 and n_2 will be connected by $P_e(n_1, n_2)$, then we have the equation

$$P_e(n_1, n_2) = \left\{ \begin{array}{ll} p, & \text{for } d(n_1, n_2) \leq r \\ 0, & \text{for } d(n_1, n_2) > r \end{array} \right\} \tag{2.1}$$

for some $0 < p \leq 1$. This work also only considers networks where the agents and connections between agents are constant over time. New agents are not created or destroyed, adjacent pairs of agents will remain adjacent, and non-adjacent pairs of agents will never become adjacent. This type of network is called a *static network* [35]. One common factor exhibited by *dynamic networks* (networks where the connections between agents can change over time) is *homophily* [6]. That is, agents are more likely to become adjacent to other agents of the same state as themselves. As this work specifically focuses on static networks, these networks do not exhibit homophily. This specification does have a significant effect on the cascade dynamics, as homophily would enable stubborn agents with minority opinions to resist change, but this factor cannot come into play in our model.

The initial distribution of active agents across the network affects the cascade propagation [12]. This work focuses on those networks where the initially active agents (called *initial seeds* or sometimes just *seeds*) are confined to a small interval of the network and every agent in that region is initially active. (For simplicity, we presume that the seed region ends at $x = r$. This avoids many of the oddities of particularly small seed sizes. Assuming the seed size to extend to some $x_0 > r$ would be pointless, as the seeds to the left of $x_0 - r$ would only be adjacent to other seeds, and be incapable of having any real influence on the cascade propagation.) If the width of the seed region is larger than the radius of influence r then there will be some seeds whose only neighbors are other seeds. These agents will not have any effect on the cascade propagation, as they cannot influence any agents capable of activating. On the other hand, if the seed region is particularly small, there is a significant chance that the cascade would terminate due to initial failure to propagate, rather than due to some geography-related phenomenon. For this reason, we assume that the seed region has width exactly r.

The distribution of agents across the geography of the network is relevant. This work considers the case where the one-dimensional network can be subdivided into many small, equally sized intervals such that there is one agent in each interval. In effect, this forces the agents to be nearly perfectly evenly distributed across the network. We refer to this distribution of agents as a *nearly fixed uniform distribution*. The nearly fixed uniform distribution

is implemented by assigning each agent an identification number from 1 to N and setting the geographic position of agent n_i as $i - u_i$ where the random variables u_i are independent uniform random variables between 0 and 1. The difference between networks with independent uniformly distributed agents and networks constructed under these assumptions is illustrated in Figure 2.3. The top image shows a network of 10 nodes with edges omitted where the nodes are distributed with a nearly-fixed uniform distribution, as defined above. Notice that there is one node in each tenth of the network. Meanwhile, the bottom image shows a network of 10 nodes that are randomly uniformly distributed. Some parts of the network have a higher concentration of nodes than others. The nearly-fixed uniform nature of the distribution of agents effectively removes the possibility of the network becoming modular. That is, there will not be any heuristic way to partition the network into distinct communities. This removes the complication of a small community serving as a catalyst or buffer for the cascade to propagate to the next community [25, 26]. This assumption is motivated by a comparison of the distribution of cascade sizes in the case of this distribution of agents and the case of independent locations of agents. (Note that the nearly-fixed uniform distribution includes a correlation between agent locations. Once an agent is placed in an interval, no other agents can be placed in that same interval.) Figure 2.4 compares the distribution of cascade sizes on two networks. Both networks are embedded on a map of width 900 and the networks both have size 900. The radius of influence of each network is 20, and the response threshold distribution of each network (denoted F and defined in section 2.2) is

$$\begin{aligned} F(1) &= 0.35 \\ F(2) &= 0.35 \\ F(3) &= 1 \end{aligned} \tag{2.2}$$

Agents near the boundary of a network will average slightly fewer neighbors (across all possible networks under our setup) than agents further away from the boundary. We define the *expected degree* of an agent to be the average number of neighbors of that agent given its location and the locations of other agents and *nominal mean degree* to be the expected degree of agents at least one radius of influence away from the boundary of the network (averaged across all possible networks following our rules for network construction). Since two agents within one radius of influence of each other have probability p of being adjacent,

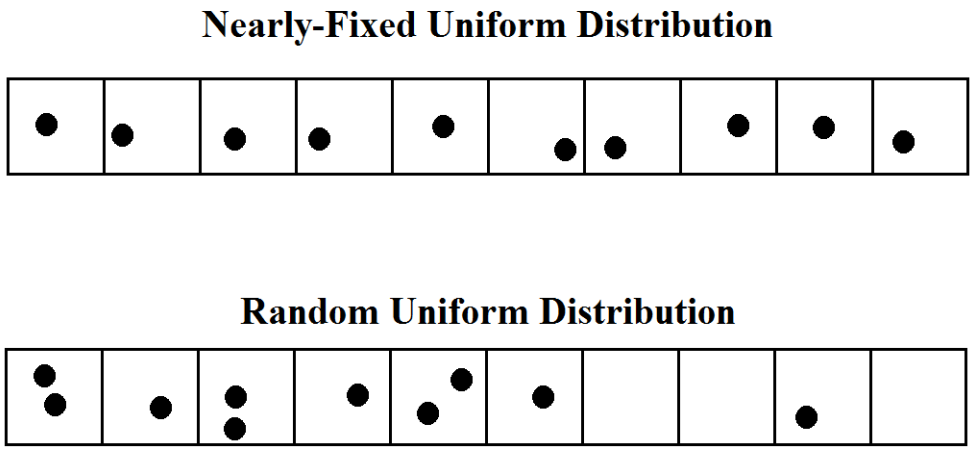

Figure 2.3: An illustration of the distinction between nearly-fixed uniform and random uniform distribution of nodes.

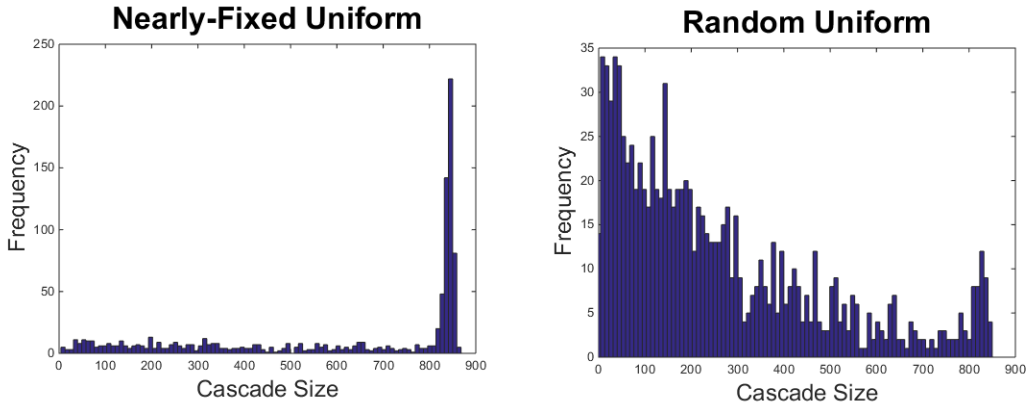

Figure 2.4: Comparison of cascade sizes on a small network of size 900, nominal mean degree 6, and radius of influence 20 using nearly-fixed uniform and random uniform distributions of agents. For both networks, the response threshold distribution was $F(1) = 0.35, F(2) = 0.35, F(3) = 1$. 1,000 cascades were simulated for each network. The termination probabilities are higher for the random uniform distribution than for the nearly-fixed uniform distribution.

then we can define the nominal mean degree to be

$$\mu = p \times (2r - 1) \tag{2.3}$$

where r is the radius of influence. (Recall that we presume the agents to be nearly evenly spaced with linear density of 1 agent per length unit. Each agent will have probability p of being adjacent to the agents r length units to its left or right, minus itself.) The true expected degree of an agent may differ from the nominal mean degree μ if the agent is sufficiently close to the boundary of the network. However, we are interested in the case where the radius of influence is small relative to the width of the network, which means that only a small fraction of agents will fit this category. For agents more than one radius of influence away from each boundary can be adjacent to the other agents within a radius of influence r of itself. This leaves $(2r - 1)$ possible neighbors. The probabilities of its being adjacent to each of these potential neighbors is $p = \frac{\mu}{(2r-1)}$, so the number of neighbors k would follow a binomial distribution with $(2r - 1)$ trials and success probability $p = \frac{\mu}{(2r-1)}$. If r is large enough and p is small, this can be approximated by a Poisson distribution with mean μ, so the likelihood p_k of the node having exactly k active neighbors is approximately

$$p_k \approx \frac{\mu^k e^{-\mu}}{k!}. \tag{2.4}$$

Because we presume the radius of influence to be small compared to the width of the whole network, we approximate the degree distribution with a single Poisson distribution with mean μ. In Figure 2.5, we compare the actual degree distribution on an instance of our toy network to the Poisson distribution with the same mean, and find that the empirical and theoretical degree distributions match well.

It has been previously found that the response threshold distribution can greatly affect the existence of partial cascades. Specifically, if the response threshold distribution is sufficiently homogeneous, with most of the agents having similar response thresholds, the cascade will be more binary, either close to the number of initial seeds or the size of the entire network [17, 20, 37]. We deliberately choose a response threshold distribution that is homogeneous enough that an Erdos-Renyi network with this response threshold distribution will have either a nearly-complete cascade or a very small cascade, as shown in the left panel of Figure 2.6. When using the same response threshold as was used in Figure 2.4, all cascades

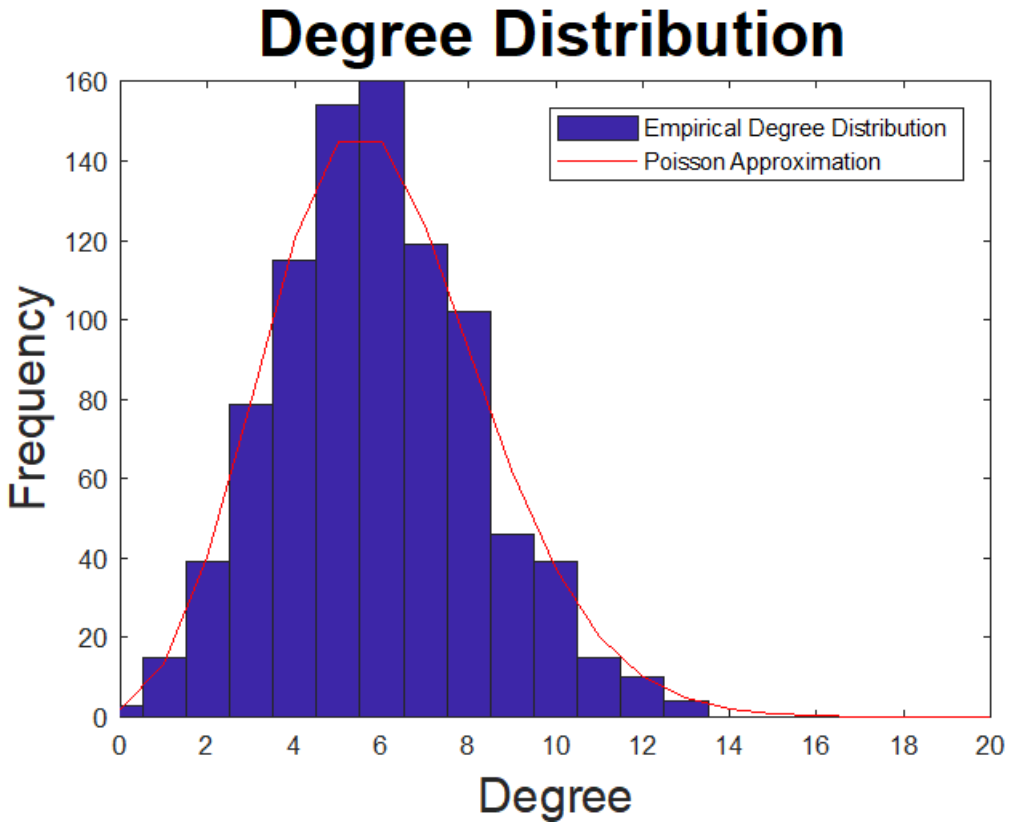

Figure 2.5: Comparison of the empirical degree distribution (blue) with the Poisson distribution of the same nominal mean μ (red).

are large.

The only difference between the network models used in Figure 2.4 is that one family has a nearly fixed uniform distribution of agents while the other has each agent independently located of each other agent and uniformly distributed across the map. Cascades terminate much more quickly on the networks with randomly uniformly distributed agents. We would like to focus on the effect of the response threshold distribution on cascade size, not the effect of relative locations of agents on cascade size. The strong effect of agent distribution on cascade size can mask the effect of response threshold distribution, so we rely on the nearly fixed uniform distribution. While the scope of this work places significant restrictions on the physical phenomena to which this method can be applied, this method may pave the way for analysis of a wider class of networks.

It is interesting to note that the permanent activation property dictates that the number of active agents is a nondecreasing function of time [9, 11, 12, 34, 35]. The number

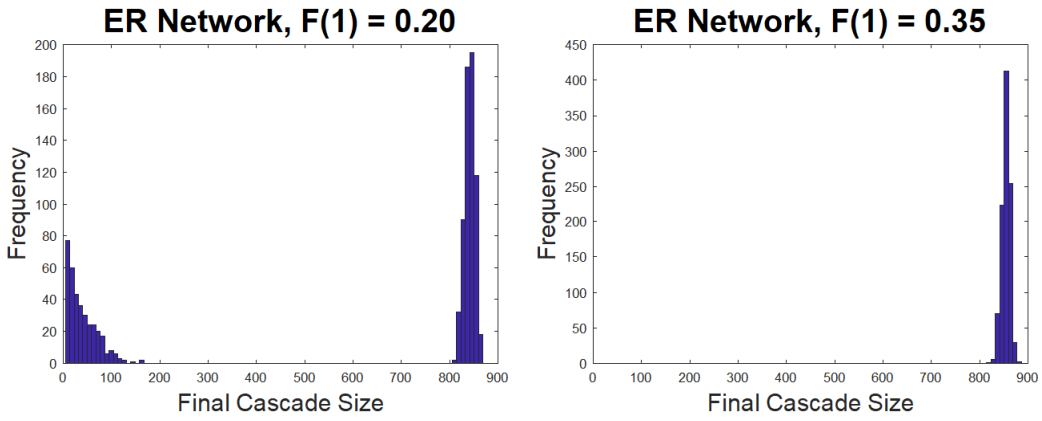

Figure 2.6: We simulate cascades on 1,000 Erdos-Renyi networks with mean degree 6, size 900, and 5 initial seeds. When use a response threshold distribution F(1) = 0.2, F(2) = 0.2, F(3) = 1, the cascades are either very large or very small. When we response threshold distribution F(1) = 0.35, F(2) = 0.35, F(3) = 1, the cascades are all very large.

of active agents is also bounded from above by the total number of agents of the network, denoted N and called the size of the network. Because the number of active agents is a non-decreasing function bounded from above, it must converge. Therefore, it is mathematically meaningful to discuss the final cascade size. Having discussed the above context, we can make the following problem statement:

"Consider a given static network of size N, following the Centola-Macy model with likelihood $P_e(n_1, n_2)$ of agents n_1 and n_2 being adjacent. Assume that $P_e(n_1, n_2)$ is dependent on a single geographic coordinate and follows the formula $P_e(n_1, n_2) = \left\{ \begin{array}{ll} p, & \text{for } d(n_1, n_2) \leq r \\ 0, & \text{for } d(n_1, n_2) > r \end{array} \right\}$ for some $0 < p \leq 1$, where $d(n_1, n_2)$ represents the distance between agents n_1 and n_2 and the probabilities of distinct pairs of agents being adjacent are otherwise independent. Suppose that the agents are distributed such that the one-dimensional network can be divided into N equally-sized segments each with one agent in it, and where the response thresholds for activation of the individual agents are independent and identically distributed with cumulative distribution function F. Assume that the initial seeds occupy some small interval of the network. Find the probability that the final number of active agents will be less than some number z."

CHAPTER 3
HISTORICAL CONTEXT AND SHORTCOMINGS OF PREVIOUS APPROACHES

Chapter 2 discussed simplifying assumptions on the rules for propagation of the cascade across the network. Even when the rules are specified, the structure of the network affects the extent to which the cascade will propagate. Previous work has studied the simplifying assumption that any pair of agents is as likely to be connected as any other pair of agents [9, 12]. Under this assumption, we only need two network statistics. The first is the fraction of agents which require only one active neighbor to activate. The second statistic is the average number of neighbors each agent in a realization of the network has. As we show in Section 3.1, these statistics are enough to predict whether the number of agents which activate on each time step increases or decreases over time. This leads to a single *cascade condition* describing when large cascades will occur and a corresponding coarse prediction of the final cascade size.

After determining whether the cascade is large or small, further analysis is conducted to determine how large the cascade is. This requires the full probability distributions of the degrees of the agents and their response thresholds. (Because we presume these values to be independent, we treat these as two separate one-variable distributions rather than a single bivariate distribution.) Section 3.2 describes this approximation in more detail and shows how well these assumptions compare to numerical data. On networks where any pair of agents is as likely to be connected as any other par of agents, this approximation works remarkably well. However, when the network structure is heavily based on geography, the approximation is less accurate. The main focus of this work is on modifying the approach to these geographic networks and these improvements are discussed starting in Chapter 4

3.1 Branching Process Approximation

Suppose that we eliminate any effect of geography on the construction of the network. That is, any two agents can be connected with some probability p, and the connections between nodes are independent. The resulting network is an Erdos-Renyi network. The

cascade problem has been heavily studied on Erdos-Renyi networks, and a few rather simple and accurate results have been found [9, 12]. These approaches assume that $p \ll 1$, meaning that the number of neighbors of a given agent is, in all likelihood, only a small fraction of the total number of agents in the network. Such a network is said to be *sparse*. This is hardly a severe limitation, as many real-world networks obey this property of sparseness.

It has been shown that on Erdos-Renyi networks, the final cascade size is consistently close to the number of initial seeds N_0 or close to the total size of the network N [9, 12]. There is a very simple procedure which can predict whether the final cascade size z will be large (close to N) or small (close to N_0). Rather than always discussing the total number of agents with a given property, it is more convenient to sometimes discuss the fraction of agents with that property. For this reason, we define the quantities ρ and ρ_0. ρ is the final fraction of active agents and can be calculated

$$\rho = \frac{z}{N} \tag{3.1}$$

and ρ_0 is the fraction of agents which are seeds and is calculated

$$\rho_0 = \frac{N_0}{N} \tag{3.2}$$

The underlying assumption is that the evolution of the cascade can be modeled as a branching process [9, 12, 31, 35]. The properties of a branching process are the following:

1) There is some initial number of active agents N_0 at the start of the process.

2) At each of a specified sequence of times, $\tau = (1, 2, 3...)$ active agents can cause other agents to activate. At any intermediate times, inactive agents do not activate.

3) Once active, an agent never deactivates.

4) If an agent became active at time τ, it can cause inactive agents to become active at time $\tau + 1$. At any other times, it cannot cause any agents to activate.

5) Initial seeds can cause other agents to become active at time $\tau = 1$ and cannot cause other agents to become active at any other time.

6) Suppose that, in accordance with rules 1) through 5), an agent is capable of activating some number of other agents at time τ. The likelihood of this agent activating exactly u other agents is some a_u. The probability a_u is the same for all agents and for all activation times τ.

7) For two active agents n_1 and n_2, the number of agents activated by n_1 is independent of the number of agents activated by n_2. The number of agents activated by a given agent is called the number of *offspring* of that agent. If agent n_i is the offspring of agent n_j, then agent n_j is said to be the *parent* of agent n_i.

Figure 3.1 shows an example of the first two time steps of an instance of a branching process. Here, there are 3 initial seeds, and the values of a_u are $a_0 = 0.1, a_1 = 0.5, a_2 = 0.4$ and $a_n = 0$ for $n > 2$. That is to say, each active agent will have 0 offspring with probability 0.1, 1 offspring with probability 0.5, 2 offspring with probability 0.4, and can never have more than 2 offspring. Unfortunately, because the visualization can only represent a single instance of the branching process, it is difficult to depict the fact that the numbers of offspring of two distinct agents are independent and identically distributed. The black circles represent the active agents. The figure is divided into three panels. The initial seeds are in the leftmost panel. The circles in the center panel represent the offspring of the initial seeds, and those in the rightmost panel represent the offspring of the agents in the center panel. Notice the presence of line segments connecting some of the circles to each other. These line segments indicate that one of the agents is the offspring of the other. Note that these line segments only connect nodes separated by exactly one panel. This is because an agent activated at τ can only activate other agents at exactly $\tau + 1$. Also, note that, while it is possible for an agent to have more than or fewer than one line segment leading to the next panel, it always has exactly one line segment coming from the previous panel. This is because each agent is the offspring of exactly one other agent, with the exception of the seeds, which are the offspring of no agents. It should be noted that while the image looks similar to the graph of a network, it is not intended to represent any underlying network structure. This distinction is important as the upcoming results consider a branching process on a network, and the adjacent agents do not necessarily have a parent-offspring relation. More specifically, if an agent n_1 were to be adjacent to another agent n_2, but neither n_1 nor n_2 caused the other to activate, there would be no line between n_1 and n_2 in Figure 3.1.

To relate a branching process to a cascade on a network following the Centola-Macy model defined in Section 2.2, we make some assumptions and approximations based on the sparse nature of the network. First, note that the seeds of the cascade can be equated to the seeds of the branching process. An agent will activate at $\tau = 1$ if the number of seeds it has as neighbors exceeds its response threshold. Because the network is sparse and the

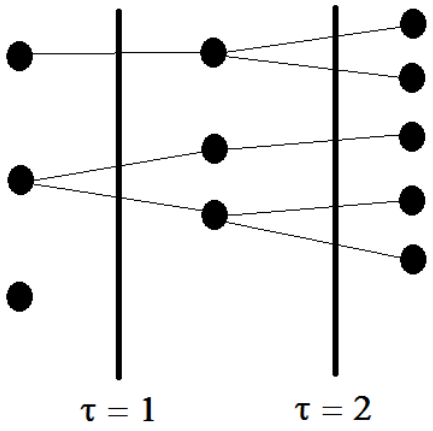

Figure 3.1: A visual representation of the first two time steps of a branching process.

probability of any two agents being adjacent is some $p \ll 1$, the likelihood of an agent being adjacent to two or more seeds is $O(p^2)$, while the probability of an agent being the neighbor of exactly one seed is $O(p)$. For this reason, we ignore the likelihood of an agent being adjacent to more than one seed. The number of neighbors of a given agent is called that agent's *degree*. The average number of neighbors of the agents across all possible instances of a network is denoted μ and is called the *nominal mean degree* of the agents on the network. Because each agent has $N-1$ other agents, each of which could be one of its neighbors, we get the approximation

$$\mu = (N-1) \times p \approx N \times p \tag{3.3}$$

If we were to construct multiple networks using the same rules, there may be some variation in the average degree of the agents in the different instances of the network. For large N, we can use the law of large numbers and would expect the approximation in (3.3) to be accurate. While we may not know the exact number of agents which are adjacent to any of the N_0 seeds, we can estimate the average number of such agents by assuming that each of the N_0 seeds is adjacent to an average of μ other agents, so that the total number of agents adjacent to a seed is approximately $N_0 \times (\mu + O(p))$. (These agents will activate if they have a response threshold of 1.) This will occur with probability $F(1) - F(0)$ where F is the cumulative distribution function of the response thresholds. The expected number of agents

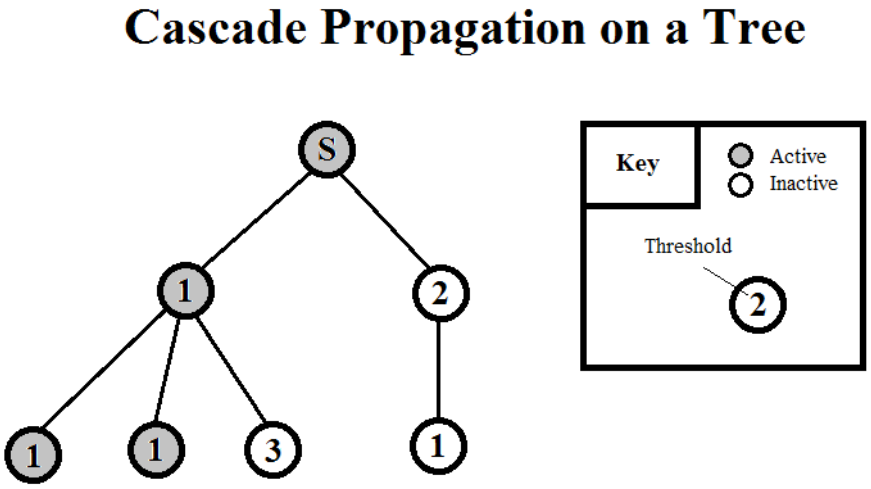

Figure 3.2: Propagation of a cascade on a small tree starting with the node marked "S" as the seed.

that would activate at $\tau = 1$ is approximately $N_0 \times \mu \times (F(1) - F(0))$. Figure 3.2 illustrates this relationship. The cascade starts with the seed at the top of the figure (marked with an S) and spreads to its neighbors with response threshold 1. Each active node, in turn, activates its neighbors on the next level down if those neighbors have response threshold 1. While the cascade is still small, we assume that none of the neighbors of a recently activated agent are adjacent to any other active agents. The assumption holds when there is a single seed and there is only one path between any two agents. Such a network is called a tree, and has no loops. While few real-world networks are trees, many sparse networks have the property that there are relatively few small loops. These network are said to be *locally treelike*, as the portion of the network that can be reached from a single agent by traversing only a few edges is nearly indistinguishable from a similar portion of a tree [27].

The calculation of the expected number of agents which activate during any time step after the first is similar, provided that the cascade size is still small compared to the network size, but has two additional complicating factors. If an agent activated at τ, then it must be the neighbor of an agent that activated at $\tau - 1$. (Because the network is sparse and presumed to be locally treelike, we assume that it is the neighbor of exactly one such agent.) This already-active neighbor cannot activate again, so it is ignored when calculating the number of agents that are adjacent to agents that activated at τ and are capable of activating at $\tau + 1$. The second factor is that, if two agents have different degrees, the one of higher degree

is more likely to be a neighbor of any given agent because it is the neighbor of more agents. Figure 3.1 illustrates the importance of these two factors. Node a will activate if three of its neighbors b, c, d, e, and f are active. Node a does not activate because only two of its neighbors (c and e) can activate. Node d would activate if node a were active, and node a would activate if node d were active, but since neither can activate without the other, neither activates. For this reason, when calculating the likelihood that node a has enough active neighbors to activate, we must presume that its neighbors could only activate without the assistance of node a. When calculating the likelihood that a node has degree k conditioned on its being a neighbor of node a, we must account for the fact that nodes of higher degree (such as node d) are more likely to be neighbors of node a than nodes of lower degree (such as node g). If we know that there are y_τ agents that activated at τ, we want to calculate the expected number of inactive neighbors of these y_τ agents. First, we ask ourselves the question "what is the likelihood that one of these agents has some degree k?" If there are no degree-degree correlations, then this question will have the same answer as "what is the likelihood that a uniformly randomly chosen neighbor of *any* uniformly randomly chosen agent has degree k?" Assume that we are given the degree distribution of all the agents in the network and find that the likelihood of a randomly chosen agent having degree k is some p_k. We define "adj" as the event of a given agent being adjacent to a uniformly randomly selected agent and we define "degk" as the event of a given agent having degree k. Using the definition of μ as the nominal mean degree, we get

$$P(\text{adj}) = \frac{\mu}{N-1}. \tag{3.4}$$

Meanwhile, an agent of degree k is adjacent to k of the $N-1$ other agents, giving us the formula

$$P(\text{adj}|\text{degk}) = \frac{k}{N-1}. \tag{3.5}$$

The degree distribution itself tells us that

$$P(\text{degk}) = p_k. \tag{3.6}$$

Bayes' theorem states that

$$P(\text{degk}|\text{adj}) = \frac{P(\text{adj}|\text{degk}) \times P(\text{degk})}{P(\text{adj})}. \tag{3.7}$$

This gets us

$$P(\text{degk}|\text{adj}) = \frac{k \times p_k}{\mu}. \tag{3.8}$$

We now move to the question "What is the likelihood that an agent has degree k conditioned on its being the neighbor of a given agent *that activated at time* $\tau - 1$?" We define "$\text{adj}_{\tau-1}$" to be the event of a specific agent being adjacent to another agent that activated at time $\tau - 1$. First, consider the case where $\tau = 1$, so the condition that the agent's neighbor activated at time $\tau - 1$ is equivalent to the agent's neighbor being a seed. Assuming that the seeds are uniformly selected among all the agents of the network, the degree distribution of a seed is the same as that of any agent, so we can use equations (3.5) and (3.6) to arrive at the result

$$P(\text{degk}|\text{adj}_0) = \frac{k \times p_k}{\mu}. \tag{3.9}$$

While seeds have the same degree distribution as other agents, the same cannot be said of agents that activated at a specific other time. The degree of an agent may be correlated with the timing of when its neighbors activated. If we randomly select an agent that activated at some time τ then the degree of that agent can be viewed as a function of that random selection, and thus as a random variable K. We calculate the expected degree of the agents that activated at time τ. We define "act_τ" to be the event of a given agent activating at time τ. We know that the agent in question was activated by another already active agent. This means that of its K neighbors, only $K - 1$ are inactive and could be influenced to activate. The likelihood of the agent in question being adjacent to another given agent, other than itself and the agent that activated it is

$$P(\text{adj}_\tau) = \frac{E[K|\text{act}_\tau] - 1}{N - 2}. \tag{3.10}$$

The numerator of the fraction in (3.10) calculates the average degree of agents that activated at time τ and subtracts the agent which activated the agent in question. The denominator of the fraction in (3.10) counts the total number of agents in the network

N and subtracts the agent in question and the agent which activated it. Expanding the expectation in (3.10), we get

$$P(\text{adj}_\tau) = \frac{\sum_k kP(\text{degk}|\text{act}_\tau) - 1}{N-2}. \tag{3.11}$$

For an agent to activate at time τ it must be adjacent to an agent that activated at time $\tau - 1$ and have threshold $T = 1$. This gives us

$$P(\text{degk}|\text{act}_\tau) = P(\text{degk}|\text{adj}_{\tau-1} \wedge T = 1). \tag{3.12}$$

Because we assume an agent's threshold and degree to be independent of each other conditioned on its being adjacent to an agent that activated at time $\tau - 1$, this reduces to

$$P(\text{degk}|\text{act}_\tau) = P(\text{degk}|\text{adj}_{\tau-1}), \tag{3.13}$$

so we can replace the probability in the expanded expectation of (3.11) to get

$$P(\text{adj}_\tau) = \frac{\sum_k kP(\text{degk}|\text{adj}_{\tau-1}) - 1}{N-2}. \tag{3.14}$$

Extending the Bayesian equation (3.8) to incorporate the time τ yields

$$P(\text{degk}|\text{adj}_\tau) = \frac{P(\text{adj}_\tau|\text{degk}) \times P(\text{degk})}{P(\text{adj}_\tau)}. \tag{3.15}$$

The reasoning behind (3.5) is time independent. An agent of degree k is adjacent to k of the remaining $N - 1$ agents, regardless of the time at which it activated, so

$$P(\text{adj}_\tau|\text{degk}) = \frac{k}{N-1}. \tag{3.16}$$

Plugging (3.14) and (3.16) into (3.15) gets the iterative equation

$$P(\text{degk}|\text{adj}_\tau) = \frac{(N-2) \times k \times p_k}{(N-1) \times \sum_k kP(\text{degk}|\text{adj}_{\tau-1}) - 1)}. \tag{3.17}$$

Since we know $P(\text{degk}|\text{adj}_0)$, (3.15) can be used repeatedly to find $P(\text{degk}|\text{adj}_\tau)$ for any τ. Because $\sum_k kP(\text{degk}|\text{adj}_{\tau-1})$ is just the expanded version of $E[K|\text{adj}_{\tau-1}]$, (3.15) can be simplified to

$$P(\text{degk}|\text{adj}_\tau) = \frac{(N-2) \times k \times p_k}{(N-1) \times (E[K|\text{adj}_{\tau-1}] - 1)}. \tag{3.18}$$

(3.18) can be simplified even further if the network is an Erdos-Renyi network. Note that if an agent was adjacent to an agent that activated at time $\tau - 1$ then to calculate its expected degree, we know that it was adjacent to the aforementioned agent that activated at time $\tau - 1$, was not adjacent to itself, and had a probability of $\frac{\mu}{N-1}$ of being adjacent to each of the other $N - 2$ agents. By linearity of expectations, this gives us

$$E[K|\text{adj}_{\tau-1}] = 1 + \frac{\mu(N-2)}{N-1} \approx \mu + 1. \tag{3.19}$$

Plugging (3.19) into (3.18), we get

$$P(\text{degk}|\text{adj}_\tau) = \frac{(N-2) \times k \times p_k}{(N-1) \times (\mu + 1 - 1)} \approx \frac{k \times p_k}{\mu}. \tag{3.20}$$

If there were some y_τ agents which activated at time τ then the expected number of inactive neighbors of those agents would be calculated

$$y_\tau \times (E[K|\text{adj}_\tau] - 1) \tag{3.21}$$

The expectation can be expanded to get

$$y_\tau \times \sum_k (k-1) P(\text{degk}|\text{adj}_\tau) \approx y_\tau \times \sum_k \frac{k \times p_k}{\mu}(k-1) \tag{3.22}$$

We are interested in the number of agents which those y_τ agents would activate. Each inactive neighbor of one of those agents would activate if it had a response threshold of 1. (Again, we neglect the possibility of an agent having multiple active neighbors when the cascade is small.) The expected number of these inactive neighbors of the y_τ most recently activated agents which activate on the next time step, which we denote $y_{\tau+1}$ is

$$y_{\tau+1} = y_\tau \times \sum_k \frac{k}{\mu} p_k (k-1)(F(1) - F(0)) \tag{3.23}$$

Gleeson and Cahalane [12] define a parameter, which we rename f, to be the expected ratio of $y_{\tau+1}$ to y_τ, which gives us the equation

$$f = \sum_k \frac{k}{\mu} p_k (k-1)(F(1) - F(0)) \tag{3.24}$$

The total expected number of agents that would activate in the interval $\tau \in (0, n]$ can be found by estimating the total number of seeds and then adding the expected number of agents that would activate at each time $\tau \in (1, 2, 3...n)$. There are N_0 seeds, each of which has expected degree μ. Assuming that the likelihood of any two seeds being adjacent is negligible, the expected number of neighbors of seeds is $\mu \times N_0$, each of which has probability $F(1) - F(0)$ of activating upon receiving one spike. Thus, we expect each of these seeds to activate an average of $\mu \times (F(1) - F(0))$ agents at $\tau = 1$, for an expected total of $N_0 \times \mu \times (F(1) - F(0))$ agents activating at $\tau = 1$. Invoking (3.24), we expect each of these to activate an additional f agents at $\tau = 2$, for an additional $N_0 \times \mu \times (F(1) - F(0)) \times f$ activated agents. Continuing this logic for each τ we get the following formula for the expected value of ρ_τ, the fraction of agents which are active at time τ

$$E[\rho_\tau] = \frac{(N_0 + N_0 \times \mu \times (F(1) - F(0)) \times (1 + f + f^2 + ...f^{\tau-1}))}{N} \tag{3.25}$$

for $\tau \geq 1$

Note that as $\tau \to \infty$ this sum converges for $f < 1$ and diverges for $f \geq 1$. As a result, this implies that the final cascade size is finite for $f < 1$ and is (somehow) infinite for $f \geq 1$. If $f \ll 1$, (3.25) implies that ρ is $O(\rho_0)$. If $f \geq 1$, the final cascade size cannot actually be infinite because the cascade size can never exceed the number of agents in the network. As the cascade grows to a sizable portion of the network, the likelihood of an agent being adjacent to more than one active agent becomes significant, and the formula for determining $E[\rho_{\tau+1}]$ from $E[\rho_\tau]$ becomes more complicated. Even with the assumption of sparseness, there are so many active agents on the network by this point that there is significant likelihood of a neighbor of a recently activated agent being adjacent to other active agents already. Similarly, after sufficient time has passed, this effect could happen as the result of a large loop in the graph, not just a small one, so the locally treelike assumption is no longer valid when estimating $E[\rho_{\tau+1}]$ from $E[\rho_\tau]$ for sufficiently large τ. One could make the simplifying assumption that, in this case, the cascade will grow to the full size of the network. Figure 3.4 compares this approximation to numerical results for various response threshold distributions on Erdos-Renyi networks with 3000 nodes and mean degree 6. In the

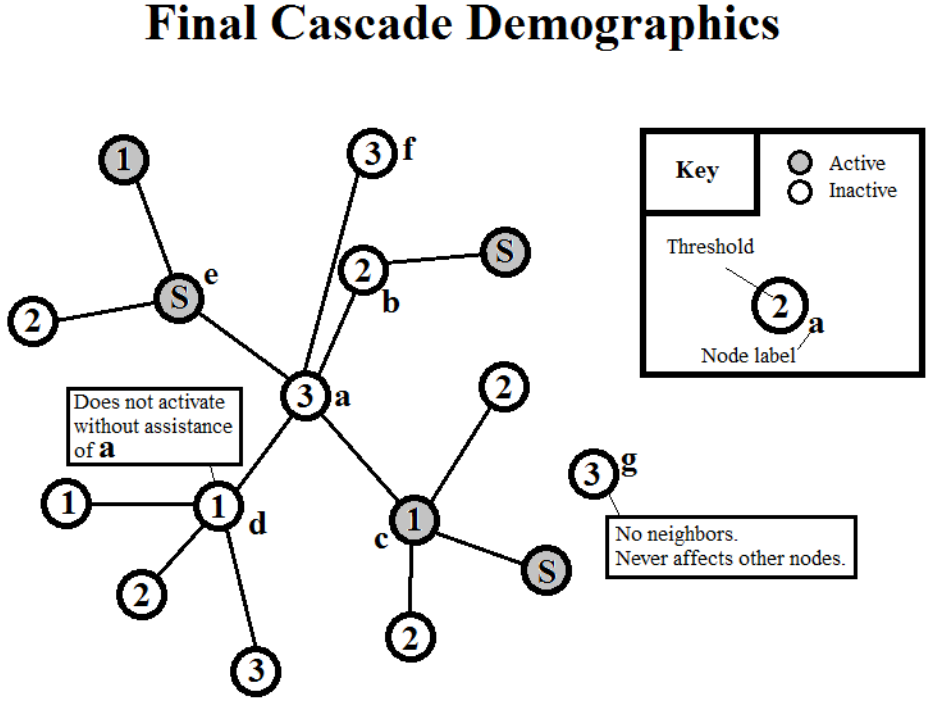

Figure 3.3: An example of a stable final cascade on a small network. The cascade starts with the three seeds (each marked with an "S") and propagates from there.

simulations plotted, $F(3) = 1$ and $F(1)$ and $F(2)$ are varied in increments of 0.05 satisfying $0 \leq F(1) \leq F(2) \leq 1$. Note the sharp transition in cascade sizes around $f = 1$. The cascades are very small for $f < 1$ and near the full network size N for $f > 1$. Gleeson and Cahalane [12] found that large cascades occur when

$$\sum_k \frac{k}{\mu} p_k (k-1)(F(1) - F(0)) > 1. \tag{3.26}$$

(3.26) is known as the *cascade condition* [9, 12, 34]. We consider the cascade condition to be a necessary and sufficient condition for a cascade to be large (on the order of N) rather than small (on the order of N_0). Previous research confirms that the cascade condition is a necessary, but not sufficient condition for large cascades [12].

The formula for f simplifies considerably if the network is an Erdos-Renyi network and p is small. We start with the formula $f = \sum_k \frac{k}{\mu} p_k (k-1)(F(1) - F(0))$. Note that because

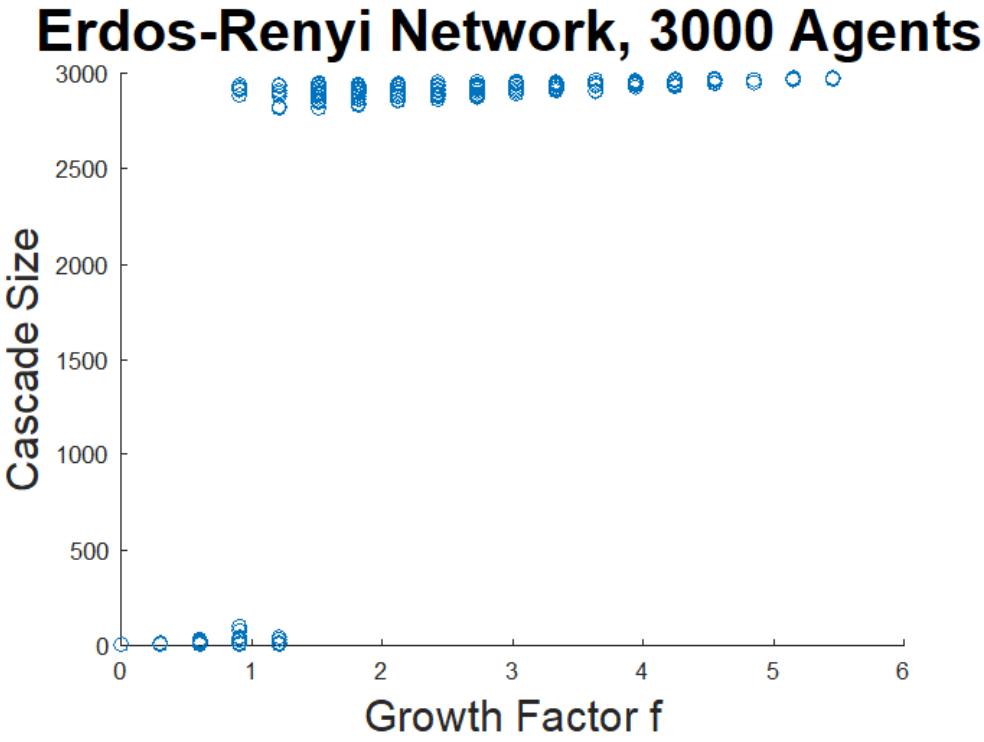

Figure 3.4: After simulating cascades on 210 Erdos-Renyi networks with varying response thresholds as indicated on page 30, we compare the final cascade sizes to their corresponding growth factors defined in equation (3.24). As earlier results show, there is a sharp transition in cascade size when the growth factor f reaches 1.

the connections between the N agents are independent, each agent could be adjacent to any of the $N-1$ other agents. Because each potential connection exists with probability p, we get the following formula for p_k:

$$p_k = \frac{(N-1)!}{k!(N-k-1)!} p^k (1-p)^{(N-k-1)} \tag{3.27}$$

In the limit of large N and small p, this approaches the well-known Poisson distribution, which has probability mass function

$$p_k = \frac{(p \times (N-1))^k}{k!} e^{-p \times (N-1)} \tag{3.28}$$

Because $\mu \equiv p \times (N-1)$, we have

$$p_k = \frac{\mu^k}{k!} e^{-\mu} \tag{3.29}$$

Putting this back into the formula for f, we get

$$f = \sum_k \frac{k}{\mu} \frac{\mu^k}{k!} e^{-\mu}(k-1)(F(1) - F(0)) = (F(1) - F(0)) \sum_k \frac{\mu^{k-1}}{(k-1)!} e^{-\mu}(k-1) \tag{3.30}$$

The sum on the right side of the equation is just the first moment of a Poisson distribution with mean μ, so it must be equal to μ. This gets us the simplification

$$f = \mu \times (F(1) - F(0)) \tag{3.31}$$

Notice that this is exactly the same as the ratio of $E[Y_1]$ to $E[Y_0]$. Thus, the cascade condition simplifies to

$$\mu \times (F(1) - F(0)) > 1 \tag{3.32}$$

for Erdos-Renyi networks. The more detailed (3.26) is needed for other locally treelike networks.

3.2 Gleeson-Cahalane Improvement to the Branching Process Approximation

This approach has an additional complication for large cascades. In the case where $f > 1$, it is not perfectly accurate to say that the cascade will envelop the whole network. While it may be true that almost every node on the network activates, the analyst studying the problem may be interested in the percentage of agents which remain inactive. For example, suppose that the cascade in question was some rapidly-spreading disease. It may be useful to know the expected number of organisms which would survive the plague. Gleeson and Cahalane previously developed a method to estimate the final cascade size more precisely [12]. The Gleeson-Cahalane model is based around simulating the cascade on a network with the same number of nodes, the same degree distribution, and whose graph is a tree. That is, the graph of the network has no cycles. (While the actual network need not be a tree, it

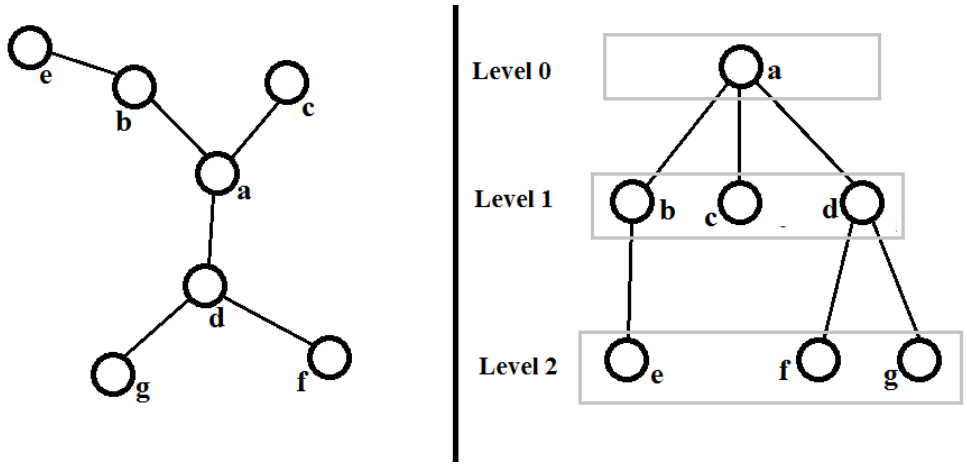

Figure 3.5: An illustration on how the nodes on a tree can be separated by level once a root has been chosen.

is locally treelike. The cascade dynamics of such a graph should be similar to those of a tree with the same size and degree distribution.) One can select any node of a tree as the root. The root is considered the origin of the tree. In a tree, there is exactly one path between any two nodes, so there is exactly one path between a node and the root. A node separated from the root by i edges is said to be of level l_i. We can then subdivide the network by level. Figure 3.5 shows a tree and another depiction of the same tree that highlights the separation of nodes by level. Note that every pair of adjacent nodes in the left panel is also adjacent in the right panel, and vice versa. Node a is the root, so it is at level 0. Nodes b, c, and d are adjacent to node a, so they are at level 1. There are paths of length 2 separating node a from nodes e, f, and g, so those three nodes are at level 2. The separation by level is evident in the right panel.

When the cascade terminates, each node is either active or inactive. In order to predict the likelihood that a given node is active, we rely on q, the likelihood that a randomly selected neighbor of an inactive node is inactive. The reason for presuming an inactive neighbor of the node comes from the fact that all non-seed nodes start out inactive. For one of these nodes n_i to activate it will need enough of its neighbors to activate while it is still inactive. During this time, each of these neighbors will need to activate despite a specific one of their neighbors n_i being inactive. The assumption that a given one of a node's neighbors is inactive is equivalent to ignoring the effects of that neighbor on the status of the node in question. Figure 3.1 shows a problem that arises if we ignore the condition of a given neighbor being

inactive in the calculation of this probability. In that figure, when we assess the likelihood of node a having enough active neighbors to reach its threshold, we must assume that those neighbors activated *before* node a would activate. Node d cannot activate before node a, so it will not activate, and neither will node a.

Gleeson and Cahalane [12] developed a formula for q through a recurrence relationship. There are two ways that a node can be active if a given one of its neighbors (namely, its parent) is inactive.

1) The node could be a seed.

2) The node could be a non-seed and have enough active neighbors, ignoring its parent, to activate.

The first case occurs with probability ρ_0. The second case requires that the node is not a seed. This condition is satisfied with probability $1 - \rho_0$. The node has some number of neighbors. Any given agent is more likely to be the parent of an agent of higher degree than the parent of an agent of lower degree. Suppose that agent a is the parent of node b. Using the same logic that led to equation (3.8) they found that the probability of agent b having degree k is

$$\frac{k}{\mu} p_k. \tag{3.33}$$

For any given agent, the probability ρ of its being active is

$$\rho = \rho_0 + (1 - \rho_0) \sum_k p_k \sum_{m=0}^{k} \mathrm{Bin}(m, q, k) F(m) \tag{3.34}$$

where q represents the probability that an agent activated without the influence of its parent. Using (3.33) as the degree distribution for the neighbors of an agent, they developed

$$q = \rho_0 + (1 - \rho_0) \sum_k \frac{k}{\mu} p_k \sum_{m=0}^{k-1} \mathrm{Bin}(m, q, k-1) F(m) \tag{3.35}$$

as an implicit formula for q, where $\mathrm{Bin}(x, p, n)$ is the probability that a binomial random variable on n trials with success probability p will have x successes and follows the formula

$$\mathrm{Bin}(x, p, n) = \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x}. \tag{3.36}$$

To find the fixed point of the implicit equation for q Gleeson and Cahalane [9] asked the similar question "What is the likelihood q_n that an agent is active, given that its parent is inactive *and it is at level* $d - n$, where d is the highest level of the tree?"

First, consider q_0. This refers to the likelihood that a leaf node at the furthest reaches of the tree would be active, given that its parent and only neighbor is inactive. This can only happen if that agent is a seed, so

$$q_0 = \rho_0 \tag{3.37}$$

At an arbitrary level of the tree, the calculation is not so simple. To activate without the assistance of its parent, an agent at this level would need a sufficient number of active children. Thus, q_n depends on the likelihood q_{n-1} of a given one of its children being active with the formula

$$q_n = \rho_0 + (1 - \rho_0) \sum_k \frac{k}{\mu} p_k \sum_{m=0}^{k-1} \mathrm{Bin}(m, q_{n-1}, k-1) F(m) \tag{3.38}$$

Assuming one of the root's neighbors was inactive, and that the root was not a seed, the root would be active with probability q_d. Assuming a large network, we approximate q_d with $\lim_{n\to\infty} q_n$.

Figure 3.7 compares this more advanced approximation developed by Gleeson and Cahalane for ρ to the numerical results for ρ. First, note that it usually predicts large values when large cascades occur, and small values when small cascades occur. Additionally, it seems to accurately estimate the number of inactive nodes when a large cascade occurs, as evidenced by the agreement between the upper-right portion of the scatterplot and the identity line.

Unfortunately, the Gleeson-Cahalane approximation has the following undesirable properties:

1) The final approximation only predicts a single final cascade size, not a probability distribution of cascade sizes. The cumulative distribution function of the resulting predicted cascade size distribution is a unit-step function.

2) The approximations predict a near-binary distribution of final cascade sizes. The final cascade size is almost always small or close to the total size of the network N.

On an Erdos-Renyi network, these issues are of minimal concern. First, the final cas-

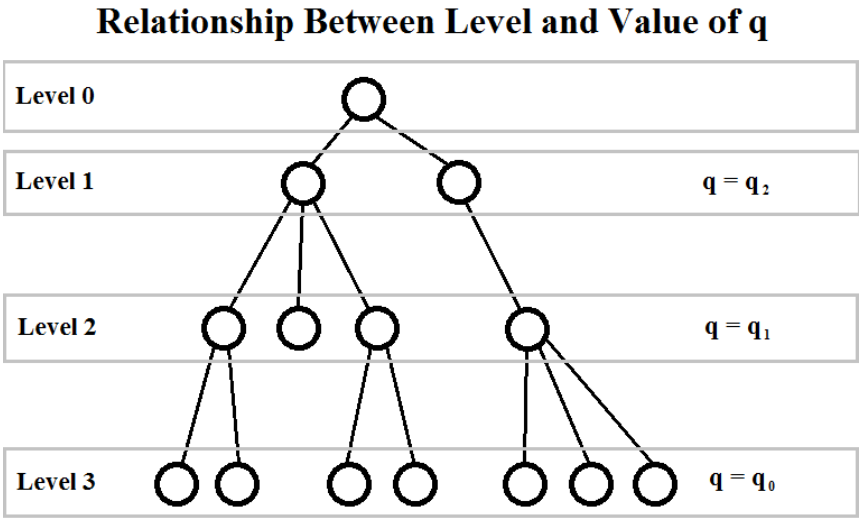

Figure 3.6: The iterative relationship between a node's level, n, and the probability q that it is active if its parent is inactive.

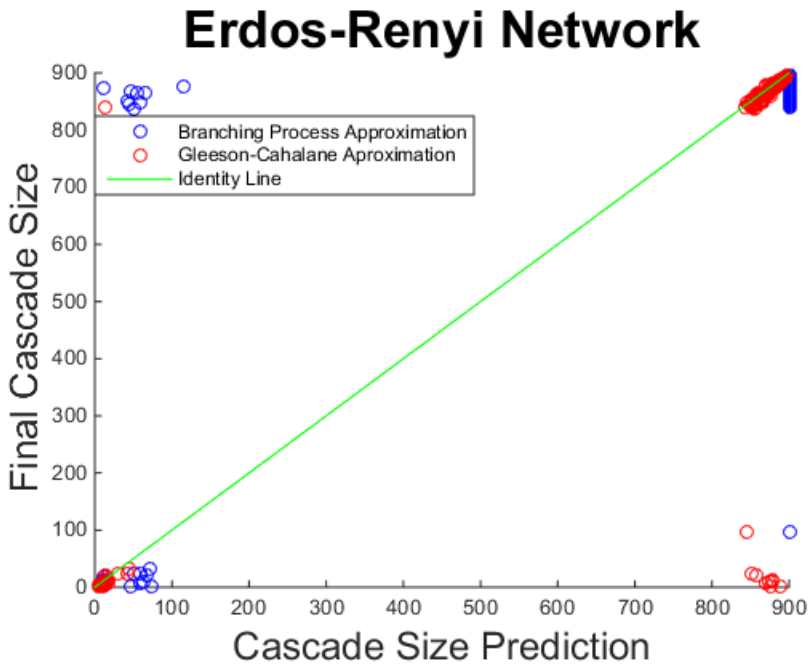

Figure 3.7: Comparison of the Gleeson-Cahalane approximation (red) and more rudimentary branching process approximation (blue) to the simulated final cascade sizes on Erdos-Renyi networks with size 900, mean degree 6, and varying response threshold distributions. $F(3) = 1$, always, $F(1)$ ranges from 0 to 1 in increments of 0.05, and $F(2)$ ranges from $F(1)$ to 1 in increments of 0.05.

cade size does display the near-binary property predicted in the Gleeson-Cahalane approximation, so this is not actually an error. Second, because of this near-binary property, the cascade size distribution can be close to a unit-step function. On a one-dimensional geographic network, these properties are no longer true. Figure 3.8 shows the cascade size distribution of a family of one-dimensional geographic networks, and Figure 3.9 shows the transition of mean cascade sizes as the response threshold distribution varies. Figure 3.8 shows the results of 10,000 cascades on networks of size 900, length 900, mean degree 6, and radius of influence 20. The response threshold distribution is $F(1) = 0.35, F(2) = 0.35, F(3) = 1$. Panel a) shows the full cascade size distribution and panel b) excludes cascades of size over 800. This makes the pattern of slightly decreasing frequency of cascades of a given size more visually apparent. Figure 3.9 shows the mean sizes of 100 cascades on networks with the same network topology as in Figure 3.8, but with varying response thresholds. $F(1)$ and $F(2)$ vary in increments of 0.05 with the restrictions that $F(1) \geq 0$, $F(2) \geq F(1)$, $F(2) < 1$. (The regions with the nonphysical trait of $F(2) > 1$ are set to 0 by default.) In all the cases plotted in Figure 3.9, $F(3) = 1$. Note that, on these one-dimensional geographic networks, the near-binary nature of the predicted mean cascade size and the inability to predict a probability distribution of cascade sizes are significant inaccuracies. While many of the final cascade sizes in Figure 3.8 congregate at the high end, the final values can realistically take any value between 0 and 900. Something about this change in the rules regarding which agents are adjacent to which other agents leads to a a smoother transition from small cascades to large cascades and leads to grater variation in the final cascade size on a network with a given degree distribution and response threshold distribution.

Most of this work is restricted to a collection of networks with specified parameters. Specifically, there is a nearly-fixed uniform distribution of agents and the networks all have mean degree 6, response threshold distribution $F(1) = 0.35, F(2) = 0.35, F(3) = 1$, size 900, width 900, and radius of influence 20. We refer to this random network model as "our toy network."

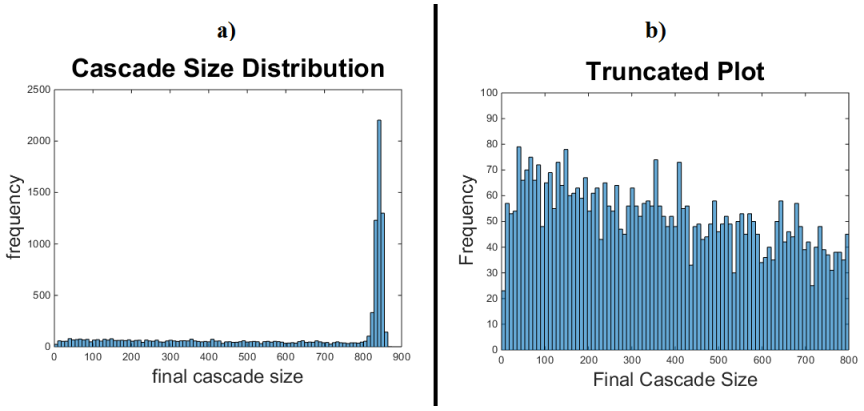

Figure 3.8: A histogram of the final sizes of 10,000 cascades on a network with a nearly-fixed uniform distribution of 900 nodes across a map of width 900. The radius of influence is 20, the mean degree is 6, and the response threshold distribution is $F(1) = 0.35, F(2) = 0.35, F(3) = 1$. Panel a) shows the full version and panel b) shows a truncated version to draw attention to non-large cascades.

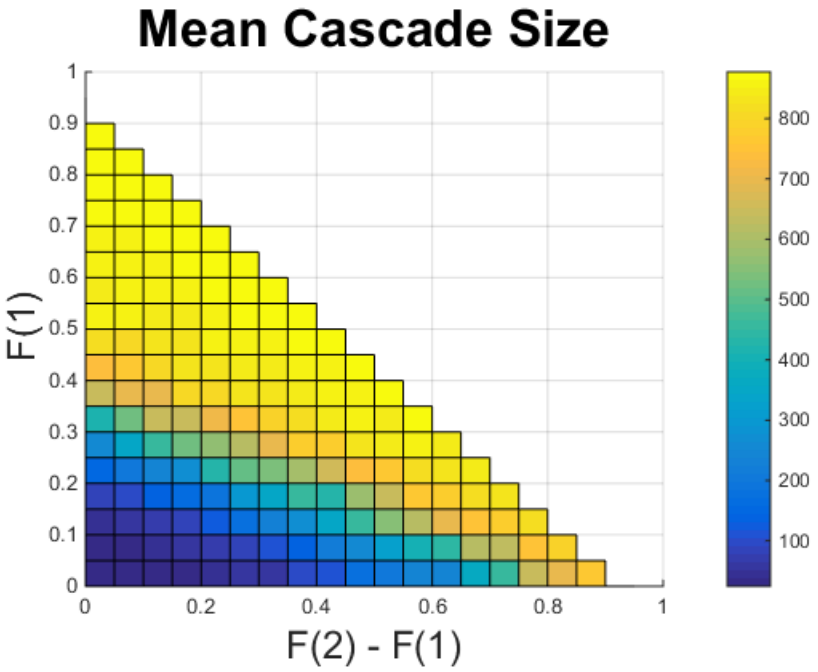

Figure 3.9: A surface plot showing the smooth transition from small cascades to large cascades as the response threshold distribution changes. The vertical axis shows the fraction of agents which require only one spike to activate and the horizontal axis shows the fraction of agents which require exactly two spikes to activate. The networks in question are similar to our toy network, but F is allowed to vary. $F(3) = 1$, always, $F(1)$ ranges from 0 to 1 in increments of 0.05, and $F(2)$ ranges from $F(1)$ to 1 in increments of 0.05. The mean cascade sizes are ensemble averages of 50 cascade sizes for each response threshold distribution.

CHAPTER 4
ASPECTS OF NETWORK TOPOLOGY WHICH AFFECT THE FINAL CASCADE SIZE DISTRIBUTION

In chapter 3, we demonstrated that our geographic networks exhibit partial cascades, where the final cascade size is neither particularly large nor particularly small. This phenomenon does not occur in Erdos-Renyi graphs, suggesting that there is some aspect of the network structure which is relevant to this qualitative difference in the cascade size distributions. Prior research [27] and conventional wisdom suggest that two relevant network statistics are the clustering coefficient and the mean intervertex path length. In this chapter, we consider the relevance of these network statistics by controlling for each one separately. We first check the cascade size distribution on a family of networks with high clustering coefficients and low intervertex path lengths. We find the same bimodal cascade size distribution found in Erdos-Renyi networks, suggesting that the intervertex path length is relevant to the change in cascade size distribution. We then conduct similar analysis on a family of networks with high intervertex path lengths and low clustering coefficient. We do not find partial cascades here, either, suggesting that the clustering coefficient is also relevant. In chapter 6 we use this information to guide our assumptions used to form a tractable approximation for the final cascade size.

It would be unreasonably time-consuming to use the entire adjacency matrix of a sizable network when estimating the final cascade size distribution. For this reason, we would like to estimate the distribution using a few simple network statistics or other summaries of the way in which nodes are connected. Intuition suggests that one relevant network statistic is the clustering coefficient C. This same intuition has led to extensive study on the effect of clustering on networks [6, 10, 13, 14, 15, 31, 32, 35, 38, 41]. The clustering coefficient describes the likelihood that two agents are adjacent given that a specific agent is adjacent to both of them. More formally, if we know that agent n_1 is adjacent to node n_2 and that agent n_2 is adjacent to agent n_3, then agent n_1 will be adjacent to agent n_3 with probability C. Figure 4.1 illustrates this concept. Node a is an initial seed. When C is non-negligible, there is a noticeable likelihood that either of its neighbor nodes b and c will have multiple active neighbors, even in the earliest stages of the cascade. Even if we were to assume that

it were impossible for any nodes from the rest of the network to send spikes to nodes b and c before those two nodes activated, there is a non-negligible likelihood that node c will be sent two spikes before it activates despite a small cascade size. Node a is a seed, so it will send a spike to node c with probability 1. Node a will also send a spike to node b, which will activate with probability $F(1)$. We examine the possibility that node b sends a spike to node c before node c activates. Consider the case where nodes b and c are adjacent, node b has response threshold 1, and node c has response threshold more than 1. At $\tau = 1$ node b will activate and node c will not. At $\tau = 2$, node b will have sent a spike to node c and node c will not already be active. The likelihood of this setup is $F(1) \times (1 - F(1)) \times C$, which is not necessarily small. The branching process approximations described in the previous section rely on the fact that if the cascade is small compared to the total network size, then the likelihood of an inactive node having multiple active neighbors is negligible. On a sparse Erdos-Renyi network (where the clustering coefficient takes the small value p, the likelihood of any two agents being adjacent) this assumption is valid. For any locally treelike network, C must be small, or there would be a significant number of three-edge loops. Since locally treelike networks cannot have many small loops, this is a contradiction. However, when C is not particularly small, the likelihood of an inactive agent having multiple active neighbors becomes significant. One would expect a geographic network to have a high clustering coefficient. Recall that the agents on such a network are restricted to be neighbors only of nearby agents. If n_1 is adjacent to n_2, and n_2 is adjacent to n_3, then the distances from n_1 to n_2 and from n_2 to n_3 must be small. By the triangle inequality, the distance between n_1 and n_3 cannot be all that large, so we would expect a greater likelihood that n_1 and n_3 are adjacent than we would if we did not know about their mutual neighbor n_2.

In 2011, Sergey Melnik and his colleagues [27] studied the effects of clustering on the accuracy of these branching process approximations. They found, surprisingly, that a high clustering coefficient could not explain the discrepancies between the approximations and the cascade size distributions of real-world clustered networks. A more relevant parameter was the mean shortest intervertex path length l between two agents. That is, for two agents n_1 and n_2, l is the smallest number of edges that can form a path from n_1 to n_2 [27].

Sparse Erdos-Renyi networks have low l and low C and do not yield partial cascades (as shown in Figure 3.4), while our toy network has high l (around 19.3, according to simulations) and high C (around 0.11) and can yield partial cascades. (On Erdos-Renyi networks with the

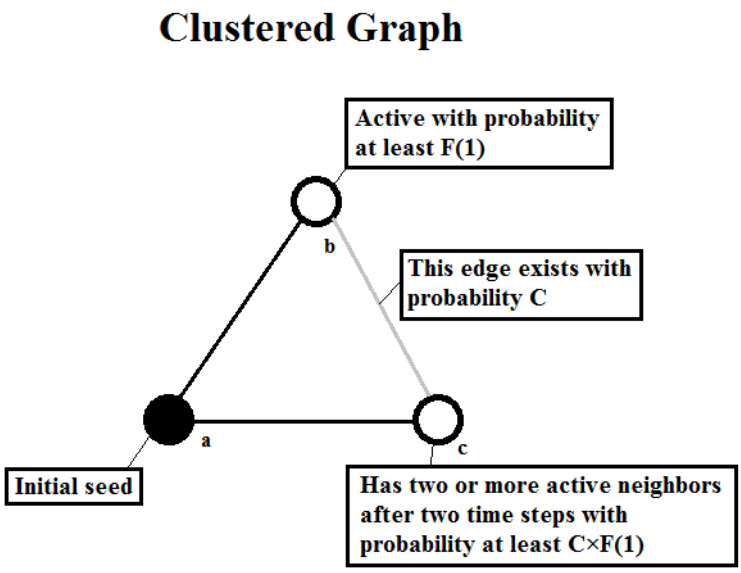

Figure 4.1: An example of how the cascade dynamics are changed when there is clustering in a graph.

same size and nominal mean degree as our toy network, $l \approx 4.1$ and $C \approx 0.007$.) We examine the effects of the clustering coefficient C and the mean intervertex path length l individually by constructing a network model with high clustering coefficient but low l, and another with a high l but low C while keeping other network statistics unchanged to the extent that we can. In each of these cases, we find the same sharp transition in final cascade sizes as predicted by the branching process approximations, suggesting that whichever modifications we make to our approximation need to reflect both an unusually high C and an unusually high l.

While Melnik et. al. [27] studied natural networks with high l and low C, none of the synthetic networks possessed those properties. To examine the case of high l and small C, we use a process similar to the L-cloning described in [8]. In the process of L-cloning, for some positive integer L, there are L identical networks and the edges are randomly rewired such that, if agents n_i and n_j are adjacent, then each copy of n_i is adjacent to exactly one randomly selected copy of n_j. The process we use, which we call "faux L-cloning", is simpler. Here, the total number of agents is multiplied by L so that there are L agents in each interval rather than only one, and the probability of any two agents being adjacent is

divided by L to maintain the same approximate nominal mean degree μ. There are still L clones of each agent in the original network, each randomly uniformly distributed within the same interval as the original agent. As no agent can be adjacent to itself, no two clones of a single agent can be adjacent either. This has minimal effect on the value of l but reduces C by a factor of approximately L. Figure 4.2 shows an example of a small network, a 3-cloned version of that network, and a faux 3-cloned version of that network. The underlying rules for generating the original network are that one node exists in each corner, each node can only be adjacent to nodes in adjacent corners (but not the corner diagonal from itself) and nodes have a $\frac{3}{4}$ chance of being adjacent to each such node. In the 3-cloned version, there are three nodes in each corner. If two nodes are adjacent in the original network, then each corresponding node in the 3-cloned network will be adjacent to exactly one of the three copies of the original neighbor. In the faux-3 cloned version, the number of nodes is tripled, but the original edges are ignored completely and new edges are put in their place in such a way that the mean degree is preserved. It is worth mentioning that a mean field approximation is equivalent to analyzing a faux ∞-cloned network. Figure 4.3 compares the final cascade size distribution on our toy network and the faux 5-cloned version of that network. The faux 5-cloned network has $l \approx 18.2$ and $C \approx 0.02$. Note that the cascade size distribution is closer to being binary in the faux cloned network. This suggests that a high l and low C will not result in the same smooth transition of mean cascade sizes and broad range of final cascade sizes as the original one-dimensional geographic network exhibits. It should be noted that the faux 5-cloned network is also a one-dimensional geographic network. Despite the two networks having very similar degree distributions, mean intervertex path lengths, response thresholds, and initial conditions, the original network has a slower evolution of the cascade (shown in Figure 6.5) and a different final cascade size distribution.

To examine the effects of increased C with l close to its value in an Erdos-Renyi network with similar size and mean degree, we use the following method of network construction. For some small integer c_i known as the *clique size*, we start with $\frac{N}{c_i}$ cliques of c_i agents each of which is adjacent to each other agent in the clique with probability 1. The mean degree of the network with only these edges is $c_i - 1$. We then add edges between pairs of agents from different cliques. Each agent would have an average of $N - c_i$ agents outside of its own clique to which it could be adjacent. Adding an additional edge between each pair of agents from different cliques with probability $\frac{\mu - c_i + 1}{N - c_i}$ would yield a network with mean

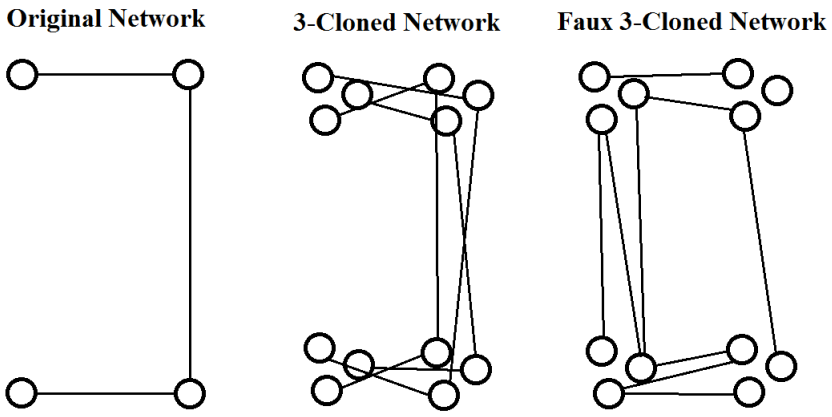

Figure 4.2: A small network (left) a 3-cloned version of that network (center) and a faux 3-cloned version of the network (right).

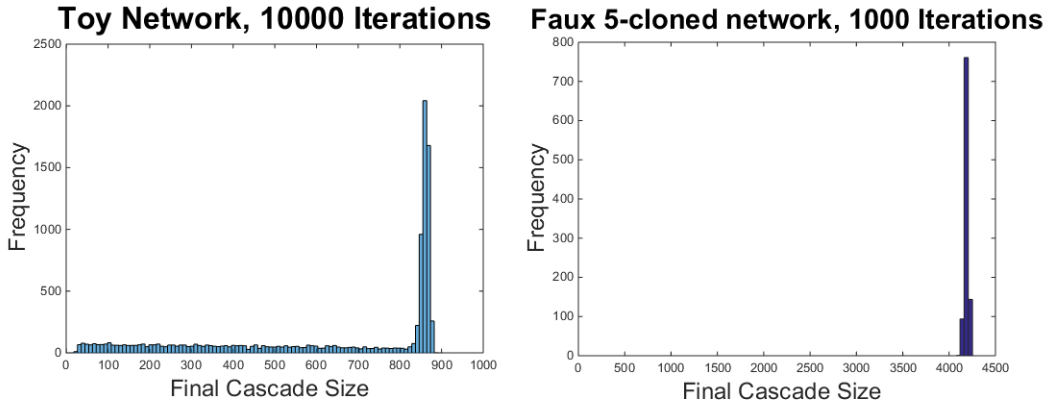

Figure 4.3: We simulate 10000 cascades on our toy network and a faux-5 cloned version of our toy network. Histograms of the final cascade sizes are plotted above. While partial cascades occur for the non-cloned network (left) they do not occur for the 5-cloned network (right).

degree approximately μ and a particularly high clustering coefficient. Figure 4.4 shows an example of the construction of such a network. Figure 4.5 compares the final cascade size distribution for a fixed response function distribution and the transition of mean final cascade size distributions as the response function is varied to the respective plots on one-dimensional geographic networks. On the clique-based network $l \approx 4.3$, similar to the value $l \approx 4.1$ for the Erdos-Renyi network, and $C \approx 0.19$, much higher than clustering coefficient of $C \approx 0.007$ on the Erdos-Renyi network. It is reasonable to assume, then, that for partial cascades to

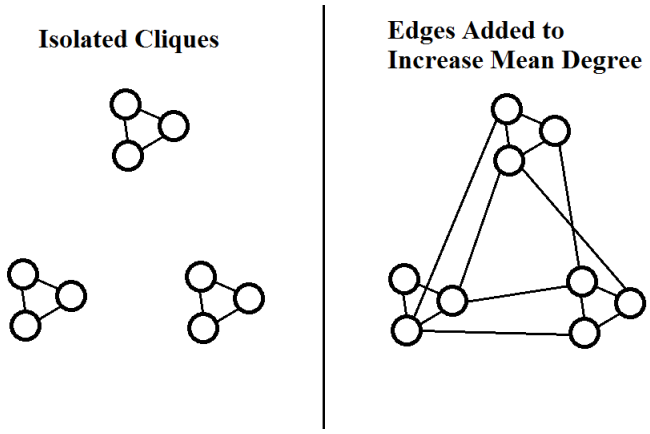

Figure 4.4: The left panel shows isolated cliques. Edges have been added to create the graph on the right panel and increase the mean degree from 2 to $3\frac{1}{3}$.

occur, both a high l and a high C are needed.

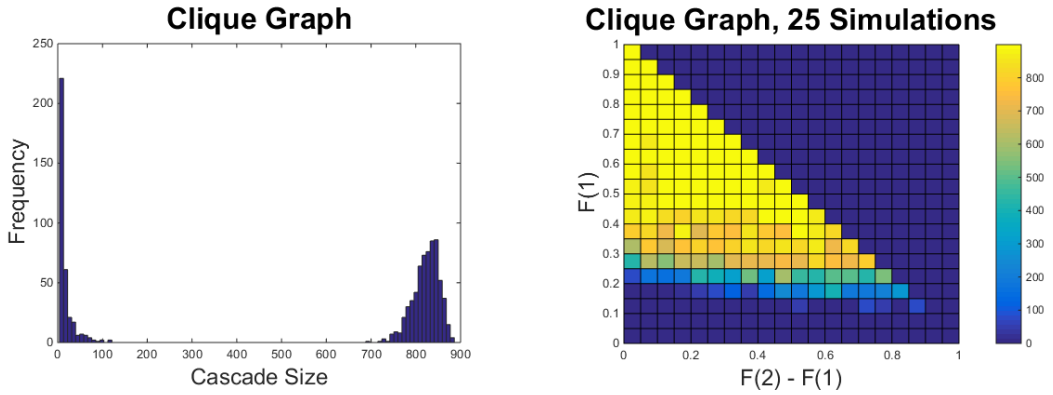

Figure 4.5: Graphical representations of the qualitative behavior of the cascade sizes on the clique-based graph. We run simulations on clique-based graphs with mean degree 6 and clustering coefficient $C = 0.3776$. The histogram of the final cascade sizes on graphs with response threshold distribution $F(1) = 0.27, F(2) = 0.27, F(3) = 1$ is plotted in the left panel. We also simulate cascades on these clique-based graphs with varying response threshold distributions. We plot the ensemble averages of 25 final cascade sizes for each response threshold distribution in the right panel. The cascades on clique graphs have extreme bimodal distributions.

CHAPTER 5
THE FUNCTIONAL FORM OF THE CDF OF THE CASCADE SIZE AND A CORRESPONDING REGRESSION-BASED APPROXIMATION FOR THE DISTRIBUTION

In this chapter, we propose a set of assumptions about the formula for the CDF of the final cascade size. We assume that the cascade will propagate across the network at a constant average rate until one of two things happens. Either the cascade will spontaneously terminate leaving a significant geographic region of the network unaffected or it will reach the natural saturation density of cascades on the network, which we call ρ_{sat}. This saturation density can be found with (3.34) developed by Gleeson and Cahalane in [12]. (On our toy network, the saturation density is 0.951.) With this limit in mind, we describe only the fraction ρ_{sat} of agents to be *capable of activating*. Until saturation is reached, we assume that the probability of extinction is constant over time, conditioned on the cascade not having terminated before then. Intuitively, this would explain the relevance of the intervertex path length and the clustering coefficient. If a given pair of agents were connected by a short path then there would not be many opportunities for a cascade to terminate after passing through one agent but before reaching the other. Meanwhile, high clustering would make local saturation more likely and make the cascade locally behave less like a branching process.

For a cascade to reach some size z, neither spontaneous termination nor exhaustion of available agents can occur before the cascade reaches size z. In Section 5.1 we examine the possibility of spontaneous termination. We use statistical analysis on a small subset of the network to estimate the propagation speed and termination probability. In Section 5.2 we focus on the possibility of the cascade reaching across the entire network, without leaving massive regions untouched. We compare the CDF predicted by this regression-based approach to numerical results on our toy network. While this approach is simple to conceptualize, it requires many simulations to conduct the statistical regression. We would prefer an approximation that could be used once for the entire prediction. For this reason, we devote chapters 6 and 7 to the development of an approach which avoids statistical regression.

5.1 A Regression-Based Approach to Approximating Cascade Size Distribution

Looking closely at Figure 3.8, we notice that using our toy network, the final cascade size seems to be able to take nearly any value from 0 to the full size of the network. Once the cascade has reached a given size, there is some nonzero likelihood that it will stop. To predict the distribution function of the cascade size, we would want to know how that probability evolves as the cascade size grows. Figure 5.1 shows the comparison of final cascade size to the number of time steps before the cascade terminated and the likelihood that a cascade would stop at a given time, provided it had not stopped before then. Notice the linear relationship between the final cascade size and the amount of time τ until termination for $\tau > 10$. For smaller τ, there is a slight positive concavity. Looking at the likelihood of termination of the cascade at each point in time, there is a very small probability of immediate termination, and a steady, approximately constant probability of termination, at least until 70 time steps have passed, where the size of the network becomes a limiting factor. We propose the following possible explanation of this behavior:

Because the number of initial seeds is fixed, the number of agents they activate is unlikely to be particularly high or particularly low. Normally there are two sources of variation in the number of activations over the course of a given time step. There is some variation in the number of activations over the course of the previous time step and some variation in the average number of new activations induced by each previous activation. At $\tau = 1$ this first source of variation is absent, so we would expect a particularly small variation in the number of activations at $\tau = 1$. A cascade will only terminate if the number of new activations is 0 (a particularly low number) so cascade termination is unlikely at $\tau = 1$. Similarly, with there being a low variance in the number of activations at $\tau = 1$ we can expect a low variance in the number of activations at $\tau = 2$, though not as low a variance as there was at $\tau = 1$. This is because the first source of variation, the uncertainty in the number of activations one time step prior, is lower than it would be some number of time steps afterward. After some relatively small number of time steps, apparently around 10, from Figure 5.1, the effect of the deterministic number of initial seeds is negligible. We assume that, after this initial period, the cascade propagates across the geographic network at a constant average speed. This assumption is corroborated by the linear nature of the relationship between the time until termination and the final cascade size. While there may

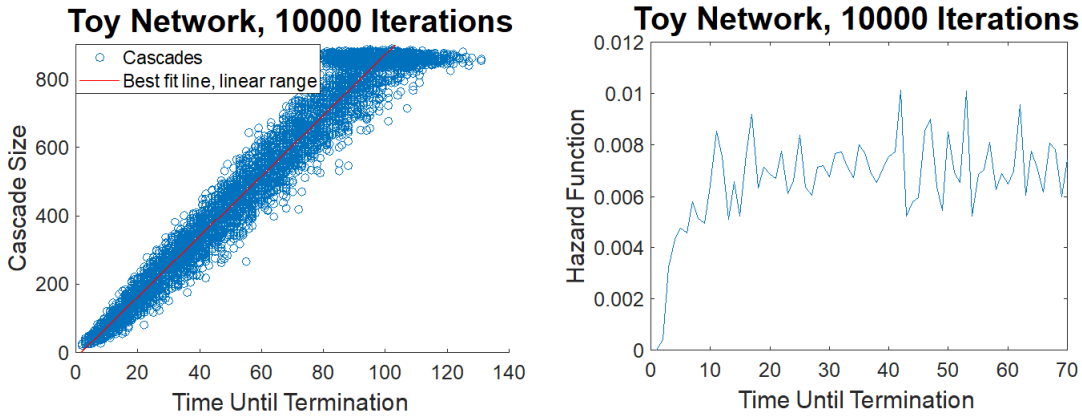

Figure 5.1: We simulate 10000 cascades on our toy network. The left panel shows the relationship between the time until termination and the final cascade size. The individual results are plotted in blue and the best-fit line of those cases where termination occurred in the linear range of $10 < \tau < 70$ plotted in red. The right panel shows the hazard function (probability of termination conditioned on termination not occurring before then) of cascade termination with respect to time. We truncate the plot on the right to exclude the increased termination probability associated with exhausting the available supply of agents, as this would be so much greater than the probability of a spontaneous termination that it would be difficult to discern the variations in the spontaneous termination probability.

be some uncertainty in the number of activations per unit time, we are only interested in the average speed. If the cascade terminates at time τ we would retrospectively expect a low number of activations at time $\tau - 1$, another low number of activations at time $\tau - 2$, though probably not as low as the number of activations at $\tau - 1$, etc. Consider two cascades, one that terminates at some time $\tau + 1$ and another which terminates at time $\tau + c$ for $c > 1$. We would expect the first cascade to be smaller than the second cascade at time τ. Suppose that the average number of new activations per unit time of an active cascade was some α. We would expect the cascade size of the second cascade at time τ to be $\alpha \times \tau$ because we would expect the cascade size to increase by α (on average) for each of the first τ time steps. In contrast, we would expect the first cascade to be slightly smaller than $\alpha \times \tau$ because we would expect the cascade to increase by *less than* α for its final few time steps. This explains why the best fit line of the termination time-to-final cascade size plot passes below the origin. The probability distribution of the number of new activations over each time step would reach a steady state, leading to a constant probability of there being

exactly 0 new activations for a given τ, evidenced by the steady nature of the termination probability across τ. Eventually, the cascade either terminates naturally or runs out of agents to activate. In this section, we restrict our attention to the case where the cascade is small enough that the system does not run out of agents to activate. We entertain the possibility of a near-complete cascade in section 5.2. Figure 5.1 suggests that the probability of termination varies between $\tau = 0$ and some $\tau = \tau^*$ and remains constant for $\tau > \tau^*$ as long as the cascade has not approached the size of the entire network yet. Assume that the probability that the cascade is still active at τ^* is some number C_1 and that the probability of termination on each time step thereafter is some number C_2. This would mean that the likelihood of its not yet terminating by $\tau^* + 1$ would be $C_1 \times (1 - C_2)$. By induction, the likelihood of the cascade not terminating by $\tau^* + n$ is $C_1 \times (1 - C_2)^n$ and the probability of its terminating before $\tau^* + n$ is $1 - C_1 \times (1 - C_2)^n$. We defined G to be the CDF of the cascade size. We now define $G_{\text{spon}}(z)$ to be the probability that a cascade *spontaneously* terminated before reaching some size z. We also define the function H to be the CDF of cascade *time* and $H_{\text{spon}}(\tau)$ to be the probability of a spontaneous termination occurring at or before time τ. We get the formula

$$H_{\text{spon}}(\tau) = 1 - C_1 \times (1 - C_2)^{\tau - \tau^*} \tag{5.1}$$

for $\tau > \tau^*$. We would like to use H_{spon} to determine G_{spon}. We presume that for $\tau > \tau^*$ the average number of new activations per time step is some number C_3. Suppose that we know that the average final cascade size of cascades that terminated at some time τ^* is some z^*. Because a cascade that terminated at time τ lasted $\tau - \tau^*$ time steps longer than one that terminated at time τ^*, we use the simplifying assumption that it would grow to a size $C_3 \times (\tau - \tau^*)$ larger than one that terminated at time τ^*, so would have size $z^* + C_3 \times (\tau - \tau^*)$. Under this simplification, the cascades that terminate before reaching size z will also terminate before time $\tau^* + \frac{z - z^*}{C_3}$. This gives us

$$G_{\text{spon}}(z) = 1 - C_1 \times (1 - C_2)^{\frac{z - z^*}{C_3}} \tag{5.2}$$

as the formula for the CDF of cascade size. This is only valid for $z > z^*$ and for z not too close to the network size N. It now remains to solve for the parameters C_1, C_2, and C_3.

Suppose that, rather than analyzing the entire network, we restrict our attention to

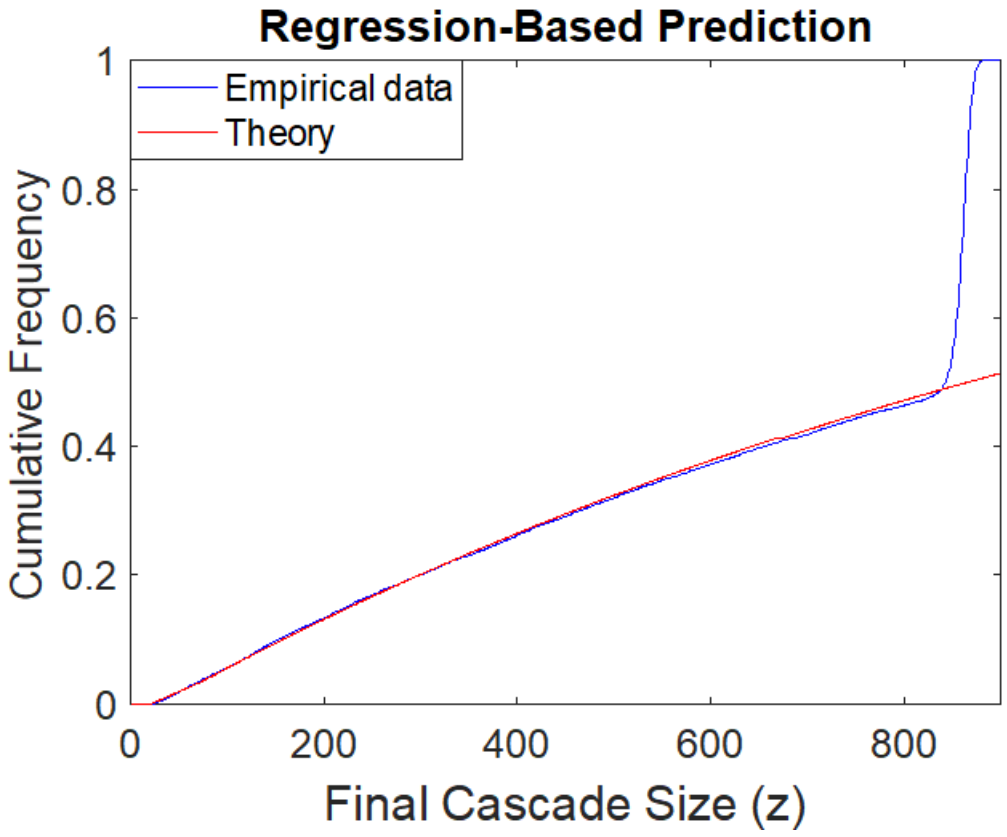

Figure 5.2: We implement our regression-based approximation for $G_{\text{spon}}(z)$ using $\tau^* = 10$ and $a = 4$. The red curve represents the resulting approximation of G. We compare this approximation to the empirical CDF approximated by 10,000 simulated cascades. This empirical CDF is plotted in blue.

a small portion of the network that includes the seed region. This small portion should be large enough that it includes any nodes that would activate before the termination probability stabilizes, as well as those that activated within some small amount of time afterward. While finding an appropriate size for this microcosm of the original network takes some guesswork, we presume an accurate guess for the moment, and discuss heuristics on how to estimate an appropriate size of the microcosm later. Assume it is conjectured that the propagation of the cascade will stabilize at some time τ^* and that the appropriate size of the microcosm is M_0. We can repeatedly run simulations of the cascade on this smaller network until some desired number s_0 of cascades terminate between τ^* and $\tau^* + a$ for some small positive integer a, and record the number of cascades s_1 that had yet to terminate after $\tau^* + a$ as well as the number of cascades s_2 that terminated before τ^*. We first estimate C_2 in equation (5.1).

Given the total number of cascades s_0 that terminated between τ^* and $\tau^* + a$ inclusive, and the number of cascades s_1 that had yet to terminate as of $\tau^* + a$, we estimate that

$$C_2 \approx 1 - (\frac{s_1}{s_0 + s_1})^{\frac{1}{a+1}}. \tag{5.3}$$

This approximation comes from the fact that of the $s_0 + s_1$ cascades that had not already terminated by $\tau^* - 1$, s_1 still had not terminated at $\tau^* + a$. We can approximate that the probability of a cascade "surviving" the $a + 1$ intermediate time steps is $\frac{s_1}{s_0+s_1}$. Because we assume that the probability of termination is the same for each of the five time steps, we assume that the probability of a cascade surviving any one of those $a + 1$ time steps is $(\frac{s_1}{s_0+s_1})^{\frac{1}{a+1}}$, so the probability of a cascade *not* surviving that time step is $1 - (\frac{s_1}{s_0+s_1})^{\frac{1}{a+1}}$.

To approximate C_1 we need to use the number s_2, the number of cascades in the simulations that terminated before τ^*. We would have the empirical estimate $\frac{s_2}{s_0+s_1+s_2}$ of the likelihood of a cascade terminating before τ^*. Because we define $H_{\text{spon}}(\tau)$ to be the probability of a cascade spontaneously terminating by time τ, we get

$$H_{\text{spon}}(\tau^*) = \frac{s_2}{s_0 + s_1 + s_2}, \tag{5.4}$$

which can be plugged into (5.1) to get

$$\frac{s_2}{s_0 + s_1 + s_2} = 1 - C_1, \tag{5.5}$$

which can be rearranged to get

$$C_1 = \frac{s_0 + s_1}{s_0 + s_1 + s_2}. \tag{5.6}$$

To estimate C_3, we find the average sizes of the cascades that terminated at each of the times $\tau \in [\tau^*, \tau^* + a]$. These average cascade sizes form a vector $\mathbf{z}$ of size $a + 1$. We can compare the vectors $\mathbf{z}$ and $[\tau^*, \tau^* + 1, ...\tau^* + a]$. Assuming the cascade had stabilized, these two vectors should be linearly correlated. We use slope of the best fit line between $[\tau^*, \tau^* + 1, ...\tau^* + a]$ and $\mathbf{z}$ as an estimate for C_3.

With this estimate, we use our toy network to verify that our choices for τ^* and M_0 were valid. If the termination speed had stabilized, the expected number of cascades that terminate at time $\tau^* + i$ would be $(s_0 + s_1) \times C_2 \times (1 - C_2)^i$. Because s_0 is known (we continue

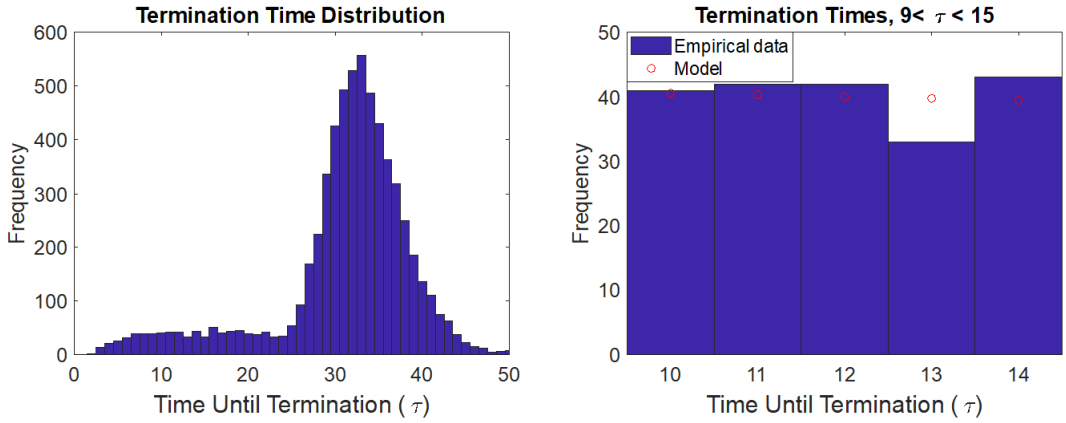

Figure 5.3: Histograms of 6007 cascade termination times (enough that there were 200 terminations in the period $10 \leq \tau \leq 14$) on the microcosm of our toy network with size $M_0 = 300$ (left) and of the termination times in $10 \leq \tau \leq 14$ compared to their theoretical values (right).

to run simulations until the desired number of cascades terminate in the appropriate window) and those s_0 cascades' termination times are distributed among $a+1$ possibilities, we can use the chi-squared test with a degrees of freedom to verify that the value of M_0 was appropriate. We set $M_0 = 300$ of the total 900 agents, $\tau^* = 10$, and $a = 4$. To vet our assumptions, we set $s_0 = 200$ (meaning that we run the simulation until 200 cascades terminate in the time interval $10 \leq \tau \leq 14$). Our simulations also resulted in $s_1 = 5807$ cascades lasting more than 14 time steps and $s_2 = 211$ cascades terminating in 9 time steps or fewer. This gets us $C_1 = \frac{6007}{6218} = 0.9661$ and $C_2 = 1 - (\frac{5807}{6007})^{0.2} = 0.0067$. Each cascade is run until it terminates, either by exhaustion or by spontaneous termination. The histogram of termination times is plotted in the left panel of Figure 5.3. Note the large peak of termination times in the interval $25 < \tau < 40$. This corresponds to exhaustion-induced terminations. Because this peak occurs well after the interval $10 \leq \tau \leq 14$ we know that our choice for M_0 was large enough that the number of exhaustion-induced cascades in the interval $10 \leq \tau \leq 14$ is negligible. Of the 200 cascades which terminate in the period $10 \leq \tau \leq 14$, we can use (5.1) to estimate the fraction of cascades that should have terminated at each time step. These theoretical values are plotted in red in Figure 5.3. The chi-square goodness of fit test gives a p-value of 0.80, suggesting that our choice for τ^* was large enough that the termination probability stabilized.

We can also verify that the propagation speed has stabilized by comparing the times

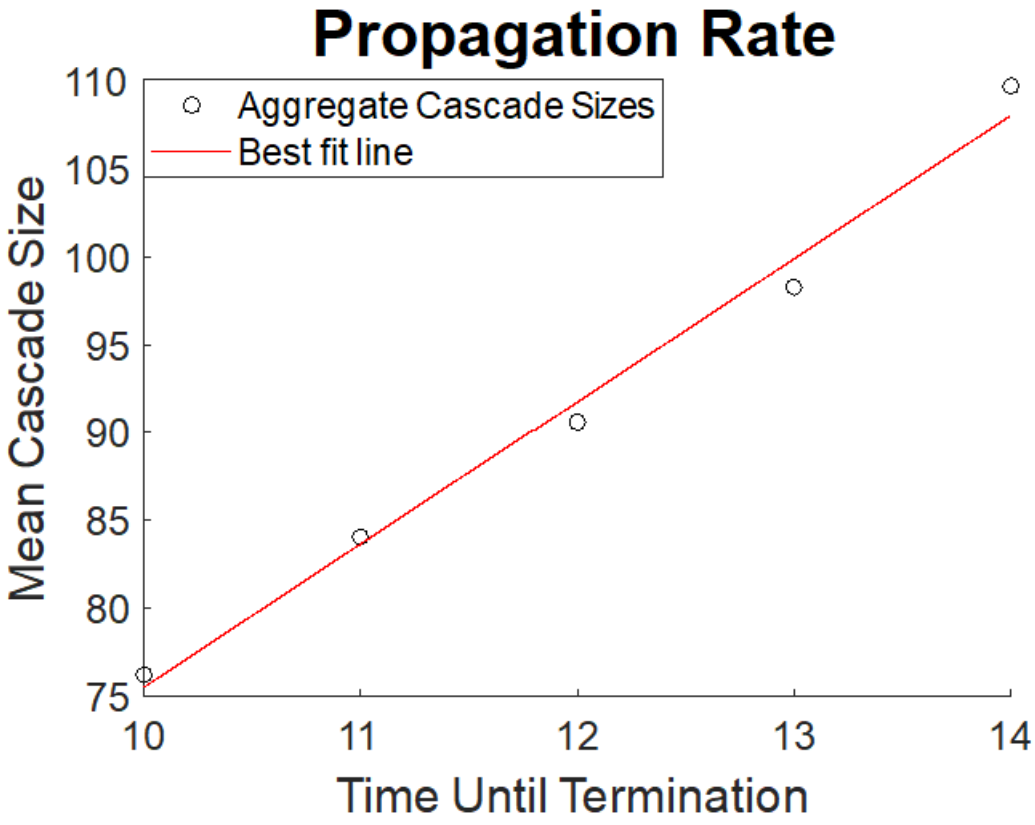

Figure 5.4: A scatterplot of the mean cascade sizes of cascades that terminated at each time in $10 \leq \tau \leq 14$ compared to its least-squares regression line.

$10 \leq \tau \leq 14$ to the average cascade sizes of the cascades that terminated at each of those times. This relationship is plotted in Figure 5.4. Because the relationship is linear we can presume that the propagation speed has stabilized. The slope of the best fit line is 8.1 and serves as an approximation for C_3 in equation (5.2).

With these estimates for C_1, C_2, and C_3 we can use (5.2) to estimate The CDF of cascade sizes due to spontaneous termination, $G_{\text{spon}}(z)$. As mentioned before, this method is only valid for cascades that have lasted long enough for the propagation speed and termination probability to stabilize. If z is small enough that we can ignore the effects of the finite network size, then $G(z) \approx G_{\text{spon}}(z)$. To approximate $G(z)$ (the CDF of the final cascade size) for z less than the cascade size z^* at time τ^*, we use a linear approximation. We know that the cascade size cannot be smaller than the number of seeds N_0, so $G(N_0) = 0$. We already assume that $G(z^*) = \frac{s_2}{s_0+s_1+s_2}$. By interpolation, we get

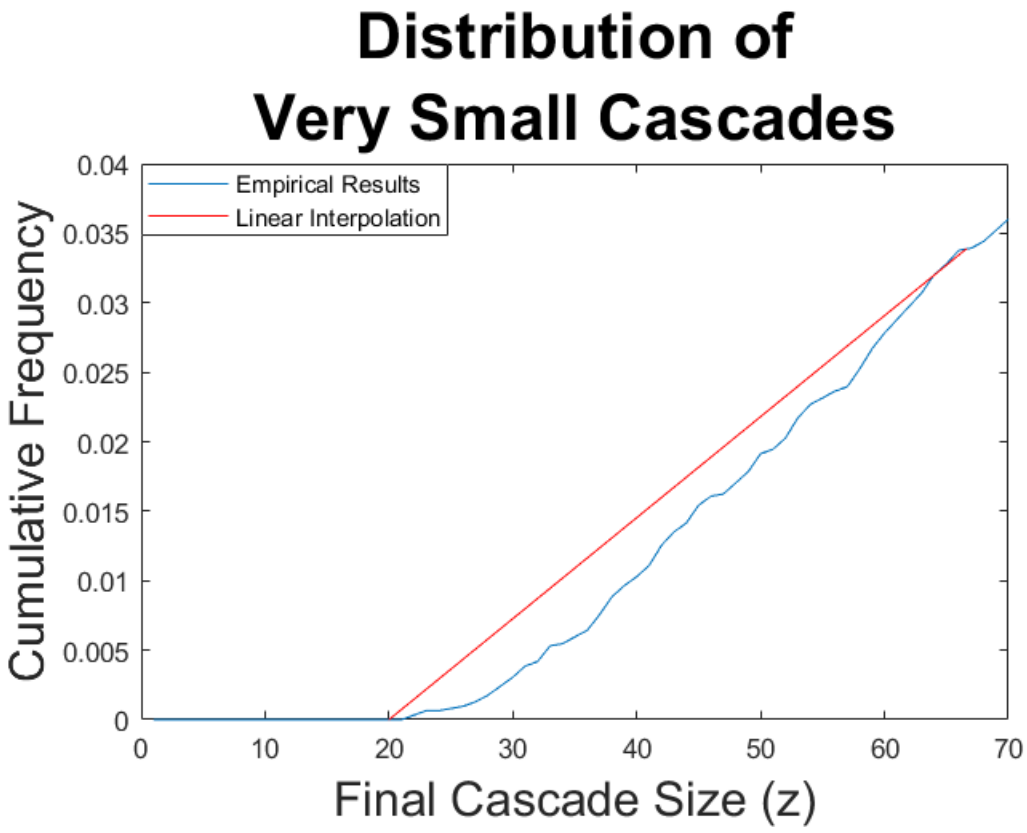

Figure 5.5: Comparison of the empirical distribution of cascades smaller than size z^* on the microcosm of our toy network (blue) to approximation using (5.7) (red).

$$G(z) = \frac{z - N_0}{z^* - N_0} \times \frac{s_2}{s_0 + s_1 + s_2} \tag{5.7}$$

for $N_0 < z < z^*$. Figure 5.5 shows this assumption of linear interpolation to be reasonably accurate.

While (5.7) and (5.2) together give us an approximation for $G(z)$ for z not close to the full network size N, they ignore the possibility of a near-complete cascade. Section 5.2 addresses the issue of large z close to the total network size N.

5.2 The Possibility of a Large Cascade

As mentioned before, there are two reasons why a cascade could terminate. The cascade could encounter a spontaneous termination or it could exhaust the supply of agents capable

of activating. For a cascade to reach some number z, there must be at least z agents capable of activating and the cascade must activate at least z agents without a spontaneous termination. We defined the function $G_{\text{spon}}(z)$ to be the CDF of the distribution of number of agents activated before spontaneous termination. We define a similar function G_{exst} to be the CDF of the distribution of the number of agents that would activate in the event of a nearly-complete cascade.

The approximation for G_{spon} that we obtain from the regression-based approximation is plotted in Figure 5.2. It has the unsettling property that $G_{\text{spon}}(N) \neq 1$. We know that the cascade cannot exceed N, so the probability of its being N or less must be 1. However, the reason that the cascade cannot exceed N has nothing to do with spontaneous termination, and is instead explained by the cascade exhausting its supply of agents capable of activating. (On our toy network, (3.34) predicts that no more than $\rho_{\text{sat}} = 95.1\%$ of agents will activate in the event of a nearly-complete cascade.) The expected number of agents capable of activating is $N \times \rho_{\text{sat}}$, as there are N agents, each of which would activate with approximate probability ρ_{sat} in the event of a full cascade. Assuming that each agent behaves independently of the others, we get the formula

$$G_{\text{exst}}(z) = \sum_{i=0}^{z} \text{Bin}(i, \rho, N) \tag{5.8}$$

This can be intractable to calculate for large N, so we approximate the binomial distribution with a Gaussian distribution with the same mean and variance. The central limit theorem allows us to do this. A measure of the error in this approximation is the magnitude of the skew of the binomial distribution, which is approximately $\frac{1}{(N\times(1-\rho))^{0.5}}$ for $\rho \approx 1$. For our toy network, $\rho = 0.951$. The skew goes to 0 for reasonably-sized networks because $N \times (1-\rho) \gg 1$. The mean of the number of agents capable of activating would be $N \times \rho$ and the variance would be $N \times \rho \times (1-\rho)$, so we get the approximation

$$G_{\text{exst}}(z) \approx \Phi(z, N \times \rho, N \times \rho \times (1-\rho)), \tag{5.9}$$

where $\Phi(x, \mu, \sigma^2)$ represents the CDF of a Gaussian distribution with mean μ and variance σ^2 as a function of x.

We plot this approximation for G_{exst} in Figure 5.6. This does get very close to 1 for $z = N$. For a cascade to reach z agents, there have to be z agents capable of activating and

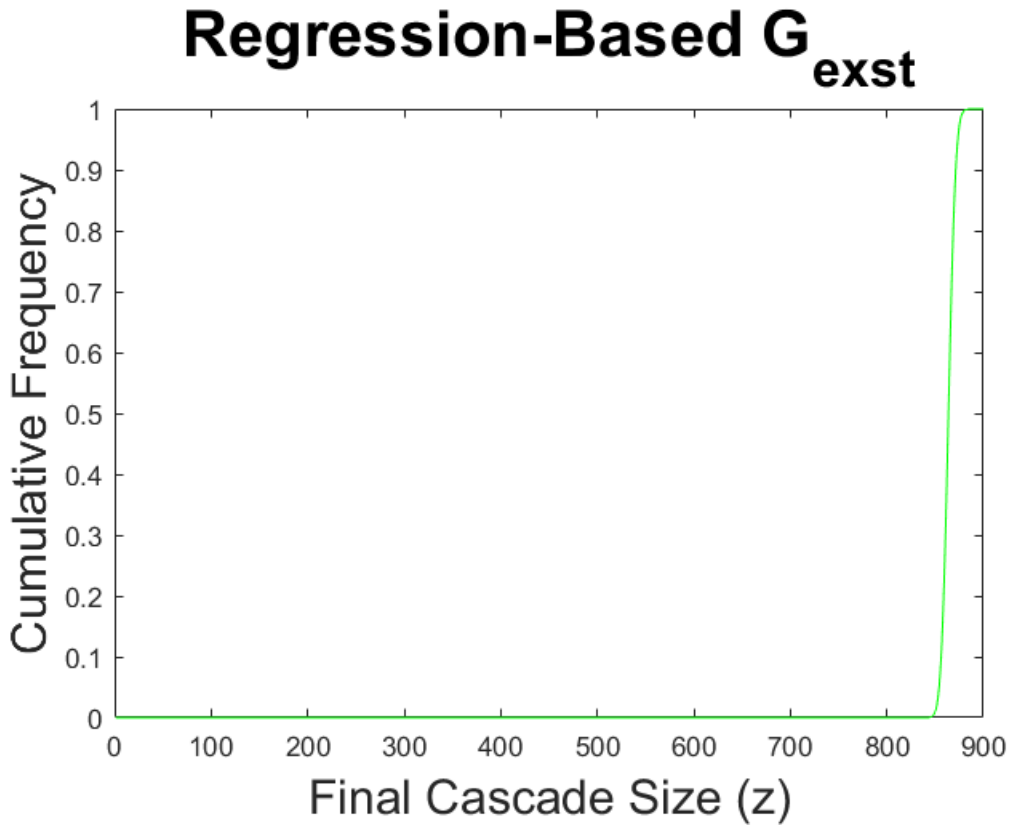

Figure 5.6: A plot of the predicted CDF G_{exst} of cascade sizes on our toy network under the assumption that the only reason for termination is exhaustion of available agents.

the cascade needs to avoid spontaneous termination for long enough to reach z agents. We assume that the events of exhaustion and spontaneous termination are independent. Thus, we get the equation

$$G(z) = \begin{cases} \frac{z-N_0}{z^*-N_0} \times \frac{s_2}{s_0+s_1+s_2} & \text{for } z < z^* \\ 1 - (1 - G_{\text{spon}}(z)) \times (1 - G_{\text{exst}}(z)) & \text{for } z \geq z^* \end{cases} \tag{5.10}$$

We plot the function G obtained by these assumptions compared to the empirical CDF of cascade sizes in Figure 5.7. The agreement is excellent.

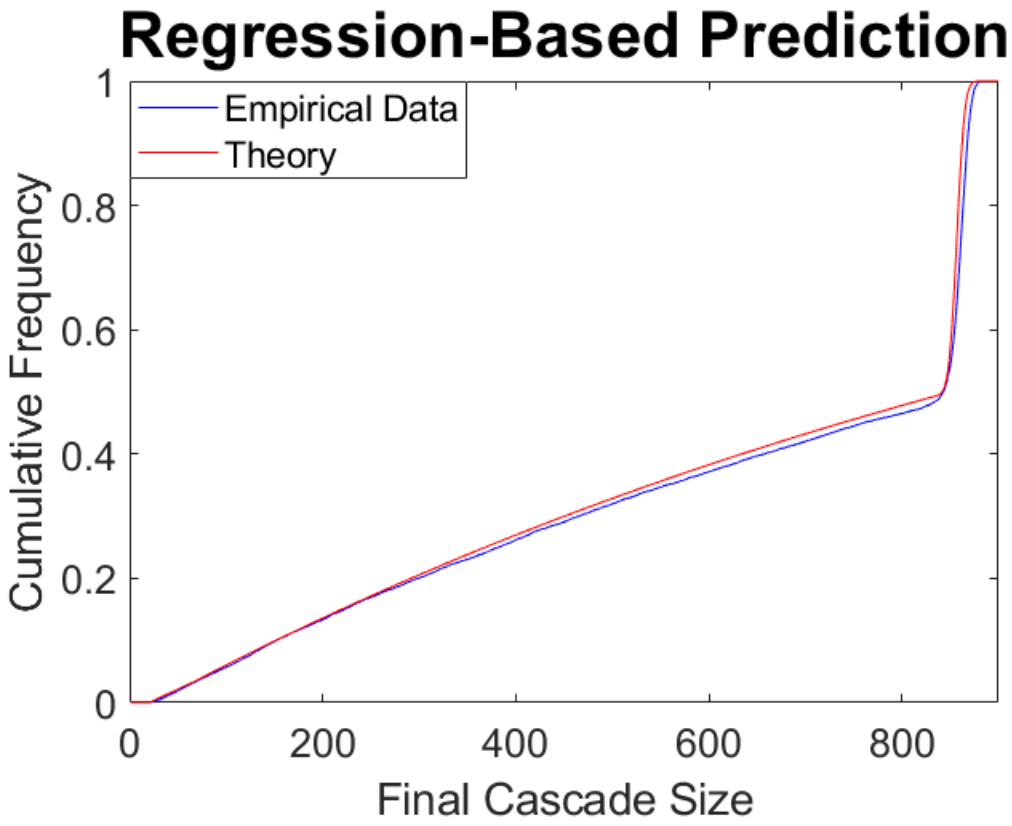

Figure 5.7: A plot of the CDF predicted if we account for both sources of cascade termination and use equation (5.10).

CHAPTER 6
A THREE-PART METHOD TO ESTIMATE CASCADE SIZE DISTRIBUTION

As indicated in Chapter 5 we assume that the cascade propagates at a constant average speed across the network and that the likelihood of termination is constant over time. We want to estimate the values of those constants, corresponding to C_2 and C_3 in equation (5.2). This would give us an estimate of the probability that the cascade would terminate at some size less than any given Z. We use a three-step process to do this. The first step is to estimate the mean propagation speed on the mean field network. We refer to the likelihood of an agent at some point x being active at time τ as the *cascade probability function* at point x and time τ, and denote it $\rho(x, \tau)$. Comparing these numbers after each time step, we can estimate the average propagation speed on this mean field network. Once the difference $\int_0^w \rho(x,\tau)dx - \int_0^w \rho(x,\tau-1)dx$ stabilizes, we use that difference as an estimate for the average propagation rate. Unfortunately, this approach does not give us the true propagation speed directly, as the propagation speed is greater on a faux L-cloned network, as will be illustrated in Figure 6.5, which compares the final cascade sizes of many simulations on our toy network to the cascade sizes on a faux-5 cloned version of that network. As the figure suggests, a cascade will propagate more quickly across a faux 5-cloned network than an uncloned network, even more quickly across a faux 25-cloned network, which can be viewed as a faux 5-cloned version of the faux 5-cloned network, etc. Thus, mean field analysis (which is equivalent to analysis of a faux ∞-cloned network) will overestimate the cascade propagation speed. As will be explained in Section 6.2, the next step of this three-part method remedies this problem.

The second step of the method accounts for the fact that the network size is finite. The underlying assumption is that once we know $Y_{\tau-1}$, the number of agents that activate at time $\tau-1$, then knowledge of $Y_{\tau-2}$, $Y_{\tau-3}$, etc will not affect the probability distribution of Y_τ. That is, we assume that the number of activations at a given time depends entirely on the number of activations during the immediately preceding time step, and not on the prior history of the number of activations. (A process with this independence of prior history is called a Markov process, or a Markov chain.) After some appropriate amount of time τ on

the averaged faux ∞-cloned network, by which point the cascade propagation has stabilized, we closely analyze the cascade probability functions $\rho(x, \tau)$ and $\rho(x, \tau - 1)$, as well as some other statistics at those times. We also assume some number y_τ agents activated at time τ. Using these parameters, we can try to estimate the likelihood that a given agent at point x that was inactive at time $\tau - 1$ had enough active neighbors at that point for it to activate by τ. (Recall that an agent will activate at time $\tau - 1$ if and only if it had enough active neighbors at $\tau - 2$.) This allows us to get the conditional probability distribution of Y_τ, the number of agents that activated at time τ, for any given $Y_{\tau-1}$. Intuitively we expect that if $Y_{\tau-1}$ is particularly large, then there will be a greater likelihood that Y_τ will be large because if there were many agents that activated over one time step then they will probably combine to activate many new agents over the following time step.

The array of the conditional probabilities of the values of Y_τ for each possible value of $Y_{\tau-1}$ (called *probabilities of transition* from $Y_{\tau-1}$ to Y_τ) is known as the *probability transition matrix* from $Y_{\tau-1}$ to Y_τ, which we denote $\mathbf{P}$. Entry (i, j) in the matrix corresponds to the conditional probability $P(Y_\tau = j - 1 | Y_{\tau-1} = i - 1)$. As a result, the i^{th} row of the probability transition matrix is the conditional probability distribution for Y_τ, assuming that $Y_{\tau-1} = i - 1$. (We use the first row for analysis of the case where $Y_{\tau-1} = 0$, which is why the i^{th} row corresponds to the case where $Y_{\tau-1} = i - 1$, not $Y_{\tau-1} = i$.) We assume that τ is large enough that the system has reached a *quasistationary state*, meaning that the distribution of Y_τ conditioned on $Y_{\tau-1} \geq 1$ is constant over time, and the probability transition matrix is constant over time. Section 6.3 describes how we estimate the probability transition matrix $\mathbf{P}$. If we can accurately estimate these transition probabilities, and if Y really is a Markov chain, then we can estimate the steady state distribution of Y_τ conditioned on $Y_{\tau-1} \geq 1$. We can then calculate the expectation of Y_τ to find the mean propagation speed and find the likelihood that $Y_\tau = 0$, which will tell us the likelihood of the cascade terminating during any given time step.

The third step accounts for the possibility of a nearly-complete cascade. In the event that the termination is caused by the system running out of available agents, then a binomial distribution is used to estimate the cascade size. This third step merges the CDF obtained by the spontaneous termination assumption with the one obtained through the nearly-complete cascade assumption. This possibility was already considered in Section 5.2, and we use the same techniques here.

Because we assume the process to be Markov, the probability transition matrix can be used to determine the probability distribution of the number of agents that activate at some time τ. Suppose that we define the row vector $\mathbf{q}_\tau$ to be the probability mass function of the number of agents that activated at time τ. If the matrix $\mathbf{P}$ was the matrix of transition probabilities from $Y_{\tau-1}$ to Y_τ, then we can calculate $\mathbf{q}_\tau$, the probability mass function of the number of agents that activated at time τ. The equation relating these quantities is

$$\mathbf{q}_\tau = \mathbf{q}_{\tau-1}\mathbf{P} \tag{6.1}$$

By the Perron-Frobenius theorem, if a Markov chain has a unique steady state, that steady state is equal to the eigenvector of the probability transition matrix corresponding to eigenvalue 1, and that eigenvalue is the one of highest magnitude. (As it turns out, if the Markov chain can transition from any state to any other state in any number of steps, then a unique steady state will exist.) Unfortunately, because our Markov chain has an absorbing state at 0, we will predict an eventual steady state of a terminated cascade. We are more interested in the *quasistationary distribution*, or the probability distribution conditioned on the cascade not terminating yet. It has been previously shown [5] that the quasistationary distribution is the left eigenvector (corresponding to the highest positive eigenvalue) of the submatrix of the probability transition matrix P with the rows and columns of all absorbing states removed. Because we assume τ is late enough that if cascade has not terminated it has reached its quasistationary distribution, we use that quasistationary distribution to find the average value of Y_τ, giving us C_3 from equation (5.2). Allowing the system to progress one more time step under the rules of the unmodified probability transition matrix $\mathbf{P}$ (this time allowing the cascade to terminate) gets us a vector $\hat{\mathbf{q}} = \mathbf{q}\ \mathbf{P}$. The first entry of $\hat{\mathbf{q}}$ corresponds to the probability that a cascade terminates at any given time conditioned on its not terminating before then, giving us C_2 from equation (5.2).

$$C_2 = \hat{\mathbf{q}}_0. \tag{6.2}$$

Suppose the expected number of new activations in quastistationary distribution was some number μ_0. We would predict that, on average, μ_0 agents would activate per unit time while the cascade was active and that it would have a termination probability $\hat{\mathbf{q}}_0$. We test to see if this approach can possibly give us an accurate CDF by using an empirically estimated

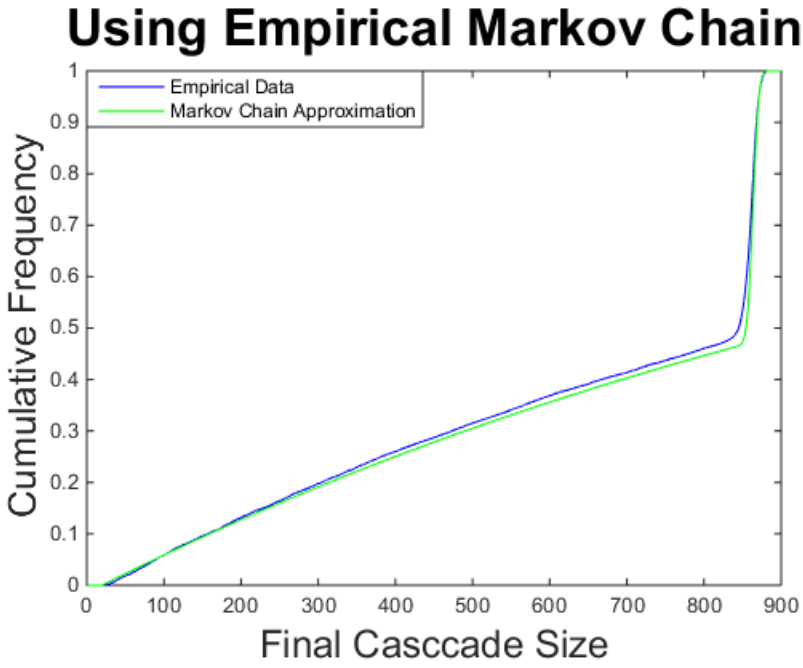

Figure 6.1: We compare the CDF of the final cascade size to the CDF that we would predict using the probability transition matrix that we empirically estimate from 30,000 simulated cascades on networks following our toy model.

probability transition matrix, taken by observing the transitions of the twentieth time step over 30,000 simulated cascades on our toy network and tabulating them into an empirical probability transition matrix. We can use this estimated probability transition matrix to estimate μ_0 and $\hat{\mathbf{q}}_0$, as we have described above, and use those parameters to estimate G, the CDF of the final cascade size. (In the case of nearly-full cascades, we make the modification described in Section 5.2.) The corresponding approximation is plotted in Figure 6.1.

We provide further explanation of how to conduct the first two parts of the approximation in the upcoming sections. We have already discussed how to account for the possibility of exhausting the supply of available agents in section 5.2.

6.1 Mean Field Analysis

It is impossible to actually construct a faux ∞-cloned version of a network. However, we can use mean field analysis to determine how a cascade on such a network would evolve over time without even declaring the locations of any nodes. Instead, we analyze the likelihood of a theoretical node at location x being active at time τ. We can do this via faux L-cloning in

the limit of large L. (Refer back to Figure 4.2 for a reminder of the faux L-cloning process.) Analysis based on looking at the deterministic averaged behavior and ignoring statistical noise is called "mean field analysis".

Of course, there are many different network realizations following our toy model, as there are numerous possibilities of which pairs of agents are adjacent. The cascade propagation along each of these networks will vary with the changes in where the edges are. As an extreme example, suppose that none of the agents in the interval from $x = 2r$ to $x = 5r$ were connected to each other. In this particular case, the cascade cannot reach any nodes with $x > 2r$. However, this is exceedingly unlikely, and becomes even less likely as L grows. In the limit as L goes to ∞, we can use the law of large numbers to predict the evolution of the cascade with increasing confidence. Suppose that for some arbitrary L, at some time τ, we knew the cascade probability function $\rho(x, \tau - 1)$ and $\rho(x, \tau - 2)$ and we wanted to estimate $\rho(x, \tau)$. The probability of a given agent in the geographic interval $x \in (i-1, i]$ being active at time τ can be found by integrating the cascade probability function over the interval, $\int_{i-1}^{i} \rho(x, \tau)dx$. For any L, the number of active agents in that geographic interval follows a binomial distribution with L trials and success probability $\int_{i-1}^{i} \rho(x, \tau)dx$. By the law of large numbers, as L grows large, the fraction of active agents in interval i in one realization of the network will converge to $\int_{i-1}^{i} \rho(x, \tau)dx$. In order to predict the likelihood that a node will activate at time τ if it was inactive at time $\tau - 1$, we need to know the likelihood that enough of its neighbors activated at time $\tau - 1$ exactly. For each inactive agent at point x with some total number of neighbors k, m of which were already active at time $\tau - 2$, we need to know the likelihood that each of its $k - m$ neighbors that were inactive at time $\tau - 2$ will activate at time $\tau - 1$ exactly. While the assumptions we use to determine this probability are not discussed until later in this section, it is a function of $\rho(x, \tau - 1)$ and $\rho(x, \tau - 2)$. If we assume that the behaviors of these neighbors of this inactive agent at x are independent, we can use a binomial distribution to construct the probability distribution of the number of active neighbors this agent will have at time $\tau - 1$ and use this distribution to figure out the likelihood of this agent activating at time τ. Doing this for all values of k and m lets us develop an estimate of $\rho(x, \tau)$ and the law of large numbers tells us that as L grows large the confidence interval about our estimate narrows. Thus, for the mean field network, we only need to figure out the evolution of the average cascade propagation across all networks that obey our parameters, and we know that $\rho(x, \tau)$ will not differ much from

that approximation at for any x and τ.

Our goal is to develop a sufficiently accurate prediction of the final cascade size using the simplest set of assumptions that we can. We make the following assumption in estimating $\rho(x,\tau)$ given $\rho(x,\tau-1)$ and $\rho(x,\tau-2)$. When determining the likelihood of a neighbor of a given inactive node activating at time τ, we weight all neighboring agents equally and consider their behaviors to be identically distributed. More precisely, consider an agent at point x_1 that was inactive at time τ with at least one neighbor that was inactive at time $\tau-2$. We declare the location of this neighbor to be x_2. We would like to estimate the likelihood that this neighbor activated at time $\tau-1$. There are $(2r-1)L$ agents within range of the agent at x_1. Of them, some were active at time $\tau-2$. The neighbor at x_2 cannot be one of those agents because we already know that it was inactive at time $\tau-2$. Of the remaining agents within range of the agent at x_1, some will activate at time $\tau-1$ (for the moment, we call this number A) while some other number (which we call B) remained inactive through time $\tau-1$. Because we weight each agent equally, the agent at x_2 is assumed to be as likely to be any specific one of the A agents that activated at time $\tau-1$ as it is to be any specific one of the B agents that did not. While the location of x_2 relative to x_1 may affect the probability that the agent at x_2 is recently activated, we ignore this effect for simplicity. Thus, we get that the likelihood of the agent at x_2 activating at time $\tau-1$ is $\frac{A}{A+B}$. Figure 6.2 illustrates this point. Here, black nodes represent agents that were active at time $\tau-2$, gray nodes represent agents that activated at time $\tau-1$, and white nodes represent agents that were inactive at time $\tau-1$. Of the four nodes within range of the node at x_1, one was active at time $\tau-2$, two activated at time $\tau-1$ exactly, and one remained inactive at time $\tau-1$. Thus, our assumption estimates that the likelihood of an agent within range of x_1 activating at time τ if it had not already activated before then is $\frac{2}{3}$. While it would be more precise to use statistics of the inactive agents at x_2 (such as the mean degree, mean threshold of agents that are still inactive, and mean number of already-active neighbors of these inactive agents) it would be too computationally intensive.

For our approximation, we keep track of the function $T_\tau(k,m,t,x)$ which measures the probability that an agent at point x will be inactive and have degree k, m neighbors that were active at time $\tau-1$, and threshold t, for $m \leq k$ and $m < t$, and is 0 elsewhere. The restriction that $m \leq k$ ensures that we exclude the physically impossible cases where agents have more active neighbors than total neighbors. The restriction that $m < t$ reflects the

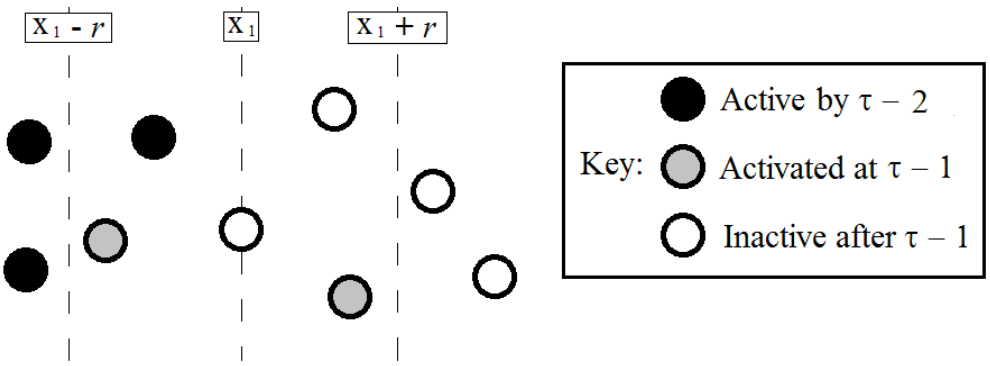

Figure 6.2: A visual representation showing the relative proximity of active vs inactive nodes to a particular node at location x_1. Of the four potential neighbors of the node at x_1, one was already active as of time $\tau - 2$, two activated at time $\tau - 1$ exactly, and one was still inactive at time $\tau - 1$. Using our current assumption that each of the four potential neighbors was as likely to be a neighbor as any other, we assume that any neighbor of the node at x_1 that was inactive at time $\tau - 2$ would have a $\frac{2}{3}$ probability of activating at time $\tau - 1$.

permanent activation principle. If $m \geq t$, the agent is already active, so there is little point in analyzing how many active neighbors it still has. Throughout our numerical estimate, we use a fourth-order tensor $\hat{T}$, defined such that $\hat{T}(k, m, t, i, \tau)$ is the approximate probability that an agent in the interval $(i-1) < x \leq i$ was inactive at time τ and had degree k, m neighbors active before time τ, and threshold t.

To begin our simulation, we initialize $\hat{T}(k, m, t, i, 0)$ to its value before the cascade even begins, when only the seeds are active. On our toy network, we set a probability $\frac{\mu}{2r-1}$ that any two sufficiently close agents are connected (in the original network, prior to cloning) to get a nominal mean degree of μ. As mentioned before, the cloning process does not affect the mean degree. However, it does affect the maximum degree. In the original network, each agent had a maximum of 40 other agents within range of itself, 19 that are definitely within range on either side, and one on either side that may be within range, unless the agent was close to the edge of the network, where the number of sufficiently close other agents would be smaller. Thus, the probability of an agent having degree greater than 40 is zero, which means that $\hat{T}(k, m, t, i, \tau)$ is zero for $k > 40$. In the faux ∞-cloned version of the network, there is no such natural maximum. This presents a problem as our tensor $\hat{T}$ needs to have finite size. For computational tractability, we set the maximum degree to 29. On a faux L-cloned version of our toy network, for large L, the degree distribution of the

agents approaches a Poisson distribution with mean 6. The probability of a Poission random variable with expected value 6 having an actual value of 30 or more is approximately 10^{-12}. As such, our assumption that no agent has degree more than 29 does not introduce any significant error and keeps our computations reasonable.

An agent is more likely to activate at some time τ if more of its neighbors activated at time $\tau - 1$. Throughout our calculations, it makes sense to keep track not only of the value of $\hat{T}$, but also of the differences between $\hat{T}$ at times $\tau - 1$ and τ. For any τ, $\hat{T}(k, m, t, i, \tau)$ depends on $\hat{T}(k, m, t, i, \tau - 1)$ and $\hat{T}(k, m, t, i, \tau - 2)$. This poses a problem for calculating the number of agents that activate at time $\tau = 1$, unless we have some way of measuring activity at time $\tau = -1$. When calculating the activity of the first time step, we would have to compare $\hat{T}(k, m, t, i, 0)$ to $\hat{T}(k, m, t, i, -1)$. In reality, there is no "-1^{st} time step," but we can interpret $\hat{T}(k, m, t, i, -1)$ to be a measure of the distributions of k, m, and t of the inactive agents before the seeds even became seeds and initiated the cascade. At that time, all agents, even the seeds, would have been inactive. The response thresholds would have already been determined, but no agents could have any active neighbors yet. Another way to explain defining $\hat{T}(k, m, t, i, -1)$ this way is to recall the conditions that will lead to an agent activating at τ. For an agent to activate at time τ and not earlier, it would need to have enough active neighbors at time $\tau - 1$ but not enough active neighbors at time $\tau - 2$. At time $\tau = 1$, though, we're only interested in the number of seeds that an agent is adjacent to, as this is the only possible source of active neighbors for that agent. Thus, we can consider any active neighbors to be new active neighbors.

To calculate $\hat{T}(k, m, t, i, -1)$ we only need to know the likelihood that a node between $x = i - 1$ and $x = i$ has response threshold t and degree k, because we know that it has 0 active neighbors. Because our model presumes that the degree and threshold of an agent are independent, we can find these two probabilities separately and multiply them together. The likelihood of an agent having response threshold t is $F(t) - F(t - 1)$. As described in Section 2.3, the probability of an agent in the interval $i - 1 < x \leq i$ having some degree k follows a Poisson distribution with mean $\mu_{\text{eff}}(i)$.

$\hat{T}(k, m, t, i, 0)$ measures the likelihood that an agent in the interval $(i - 1) < x \leq i$ will be inactive at time 0 and have degree k, response threshold t, and m active neighbors at that time. For locations i in the seed region, $\hat{T}(k, m, t, i, 0) = 0$, because these agents will not be inactive. For locations i outside the seed region, no agents activated prior to $\tau = 1$, and no

agents changed degree or response threshold, as this is a static network, so the probability that one of these agents has degree k and response threshold t is the same as it was before time 0. Agents in the seed region are of minimal interest to us, because they cannot be *induced* to activate. Agents particularly close to the opposite end of the network are also of minimal relevance at this point because if the cascade could reach that far, it would be a nearly-complete cascade. With these considerations in mind, we take $\mu_{\text{eff}} = \mu$, where μ is the nominal mean degree of the network. What changes is the likelihood that an agent with these properties will have some number m active neighbors. Figure 6.3 illustrates how we calculate this probability. Consider node 1 outside the seed region, but within one radius of influence of the seed region. If the agent is at point x, then it is a potential neighbor of any seeds between $x - r$ and r. (Recall that we presume the seed region to end at $x = r$.) With the set of potential seed neighbors of this agent occupying this interval of width $2r - x$ and the set of all potential neighbors occupying two intervals of combined width $2r - 1$, the likelihood of any one of the agent's neighbors being a seed is $\frac{2r-x}{2r-1}$. For an agent of degree k, its number of active neighbors follows a binomial distribution with k trials and success probability $\frac{2r-x}{2r-1}$. In contrast, for node 2, which is separated from the seed region by more than r, the number of seed neighbors must be 0. Taken together, these two cases give us the equation

$$\hat{T}(k, m, t, i, 0) = \left\{ \begin{array}{ll} 0 & \text{for } i \leq r \\ (F(t) - F(t-1))\frac{\mu^k e^{-\mu}}{k!}\text{Bin}(m, q(i), k) & \text{for } i > r \end{array} \right\} \tag{6.3}$$

where $q(i)$ represents the probability of the neighbor of an inactive agent in the interval $i - 1 < x \leq i$ being active (which can only happen at time 0 if that neighbor is a seed) and is estimated $q(i) \approx \max(0, \frac{2r-i+0.5}{2r-1})$. (Since we do not know an agent's exact location, but only know that it is in some particular interval $(i - 1) < x \leq i$, we have approximated x by $i - 0.5$.)

We use our approximation to estimate $\rho(x, 1)$, the cascade probability function at time 1 for the mean field network. For our toy network, the seed region covers the interval $0 \leq x \leq 20$, so $\rho(x, 1) = 1$ over this interval. Outside the seed region an agent will be active at time 1 if it had enough active neighbors at time 0, so it would need enough seed neighbors. We know that, for an agent in interval i, each of its neighbors will be active with probability approximately $q(i) = \max(0, \frac{2r-i+0.5}{2r-1})$ and that these neighbors behave nearly

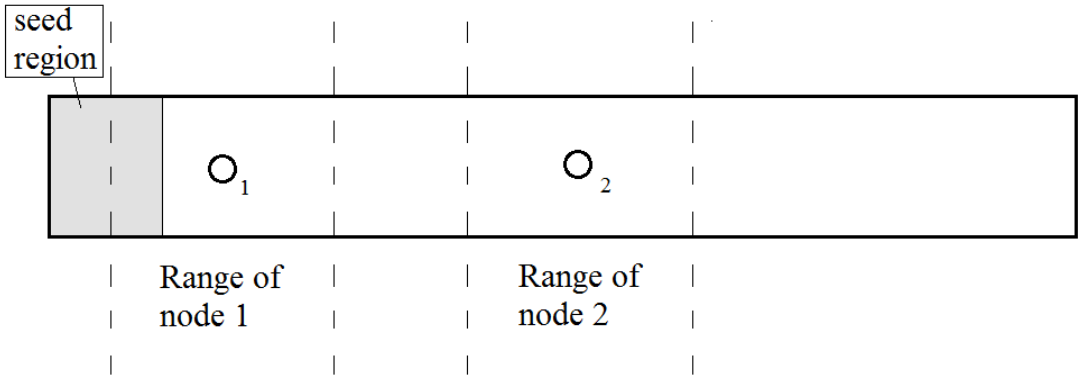

Figure 6.3: The node closer to the seed region can be adjacent to seeds, while the one further from the seed region cannot.

independently. (They do both depend weakly on the agent's position within its interval. An agent at coordinate $x = 22.01$ will average slightly fewer seed neighbors than an agent at coordinate $x = 22.99$, which causes a slight correlation between the states of neighbors of a given inactive agent. We ignore this effect.) The number of seeds to which a given agent of degree k is adjacent follows a binomial distribution on k trials and success probability $q(i)$. Accounting for all possible degrees of an agent at a given location and all possible response thresholds, we get

$$\rho(x,1) = \left\{ \begin{array}{ll} 1 & \text{for } x \leq r \\ \sum_k \sum_t \sum_{m=t}^{k} (F(t) - F(t-1)) \frac{\mu^k e^{-\mu}}{k!} \text{Bin}(m, q(\lceil x \rceil), k) & \text{for } x > r \end{array} \right\} \tag{6.4}$$

where $q(i)$ represents the likelihood that a neighbor of an agent in interval i was active at time 0, which is only possible if that agent is a seed. Because an agent will activate if its number of already active neighbors m is sufficiently high, we need to ascertain the probability of m exceeding the response threshold t. We can view the number of seed neighbors of an agent at location x as Poisson distribution with mean $\mu \times q(\lceil x \rceil)$. As such, equation (6.4) simplifies to

$$\rho(x,1) = \left\{ \begin{array}{ll} 1 & \text{for } x \leq r \\ \sum_{m=0}^{\infty} \frac{(\mu \times q(\lceil x \rceil))^m e^{-\mu \times q(\lceil x \rceil)}}{m!} F(m) & \text{for } x > r \end{array} \right\}. \tag{6.5}$$

For $\tau > 1$ the approximations of $\rho(x,\tau)$, $T_\tau(k,m,t,x)$, and $\hat{T}_\tau(k,m,t,i)$ rely on the

values of the corresponding functions from $\tau - 1$ and $\tau - 2$. For an agent at coordinate x, we want to compare the total number of agents within reach of x that were inactive at time $\tau - 2$ to the number that were inactive at time $\tau - 2$ but activated at time $\tau - 1$. The first of those quantities is $\int_{R(x)}(1 - \rho(x, \tau - 2))dx$ and the second is $\int_{R(x)} \rho(x, \tau - 1) - \rho(x, \tau - 2)dx$, where $R(x)$ is the region which an agent at x can reach. For a given x, $R(x)$ covers the interval from $x - r$ to $\lfloor x \rfloor$ and from $\lceil x \rceil$ to $x + r$. We can calculate $\bar{q}(x, \tau - 1)$, the likelihood of a neighbor of a node at coordinate x being active at time $\tau - 1$ if it was inactive at time $\tau - 2$ as

$$\bar{q}(x, \tau - 1) = \frac{\int_{R(x)} \rho(\hat{x}, \tau - 1) - \rho(\hat{x}, \tau - 2)d\hat{x}}{\int_{R(x)}(1 - \rho(\hat{x}, \tau - 2))d\hat{x}} \tag{6.6}$$

If n of an agent's $k-m$ previously inactive neighbors become active, and the agent itself does not activate, then $T_\tau(k, m, t, x)$ decreases, $T_\tau(k, m + n, t, x)$ increases, and ρ remains unchanged. One way to calculate the value of $T_\tau(k, m, t, x)$ would be to convolve $T_{\tau-1}(k, m - n, t, x)$ with $p(n|k, m, x, \tau)$, the probability that an inactive agent at point x with degree k and m active neighbors received n additional spikes at time τ. Because we presume the agent's neighbors to act independently, the distribution of n follows the binomial distribution on $k - m$ trials with success probability $\bar{q}(x, \tau)$.

$$p(n|k, m, x, \tau) = \text{Bin}(n, \bar{q}(x, \tau), k - m) \tag{6.7}$$

However, if n is high enough that the agent at x activates, it should no longer be "counted" by the function T which measures the probability of an agent at a location *being inactive* and having a given degree, number of already active neighbors, and response threshold. This gives us the formula

$$T_\tau(k, m, t, x) = \begin{cases} \sum_{n=0}^{m} T_{\tau-1}(k, m - n, t, x) \times p(n|k, m, x, \tau) & \text{for } m < t \\ 0 & \text{for } m \geq t \end{cases} \tag{6.8}$$

for $m < t$, and 0 otherwise.

If n is high enough that the agent itself does activate, then $T(k, m, t, x)$ decreases between times $\tau - 1$ and τ, $T(k, m + n, t, x)$ remains unchanged, and ρ increases. This gives

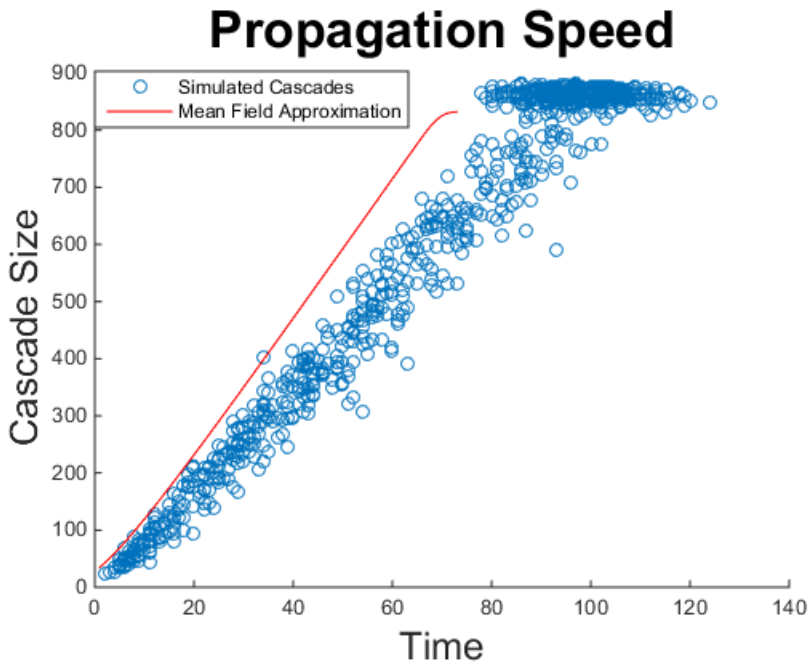

Figure 6.4: We compare the final cascade size to the time to termination using the mean field approximation of our toy network (red) to the results of 1000 simulations of our toy network (blue).

us the formula

$$\rho(x,\tau) = \rho(x,\tau-1) + \sum_{k}\sum_{t}\sum_{m=0}^{\min(k,t-1)}\sum_{n=t-m}^{k-m} T_{\tau-1}(k,m,t,x) \times p(n|k,m,x,\tau). \tag{6.9}$$

As we increase τ, we develop an estimate of the cascade's propagation across the network. The theoretical propagation across the toy network is plotted in red in Figure 6.4. The theoretical propagation is compared to the final cascade sizes of 1000 simulations, plotted in blue. Unfortunately, the mean field theory overestimates the average propagation speed. This discrepancy is addressed and corrected in Sections 6.2 and 6.3.

6.2 Discrepancy Between Cloned and Uncloned Propagation Speed

The average propagation speed predicted by our method is noticeably higher than the empirical results. In general, faux L-cloning results in a greater propagation speed for larger values of L. This phenomenon is demonstrated in Figure 6.5, which compares the final

cascade sizes and termination times for 1,000 cascades on our toy network and 1,000 cascades on a faux 5-cloned version of that network. Because there are 5 times as many agents on the faux 5-cloned network, and because the linear density of agents is 5 times greater, we scale the cascade size down by a factor of 5 to more directly compare the geographic distances that the cascades propagated. (At this point, it becomes important to distinguish between the *propagation speed* which reflects the number of number of agents which activate per unit time, and the ***geographic*** *propagation speed* which measures the distance traveled by the wave front per unit time.) It tends to take less time for the normalized cascade to reach a certain value on the faux 5-cloned network than on the original network. The discrepancy between the mean field propagation speed and the actual propagation speed on the uncloned network stems from the concavity of $E[Y_\tau|Y_{\tau-1} = y_{\tau-1}]$ as a function of $y_{\tau-1}$. As indicated in the left panel of Figure 6.9, the curve has a downward concavity. By Jensen's inequality, the statistical noise will decrease the unconditional $E[Y_\tau]$. If we were to use a random uniform distribution of agents instead, this heterogeneity of the network would have a greater effect on the difference in propagation speed, as the cascade would spend more time in the slow regions while passing through the fast regions more quickly. Because this is not the case, we do not need to concern ourselves with the linear density in the vicinity of the most recently activated agents. (This is verified by the accuracy of prediction using the empirical probability transition matrix in Figure 6.1.)

Because mean field analysis is the same as using a faux ∞-cloned network, the propagation speed predicted this way will exceed the actual value. Unfortunately, this prevents us from solving for the mean propagation speed and the likelihood of cascade separately. If the mean propagation speed on the faux ∞-cloned network were an unbiased estimator of the mean propagation speed on the original uncloned network then we could use this easily-obtained estimate and only be left with the challenge of finding the likelihood of a cascade terminating. Unfortunately, we do not have this convenience. This issue also presents a difficulty in finding the source of inaccuracies in a given prediction. Ideally, we would prefer to know whether an error could be explained by an inaccuracy in the faux ∞-cloned propagation, an inaccuracy in the construction of the probability transition matrix from an accurate propagation rule, or both. Unfortunately, we cannot examine whether or not the propagation speed is accurate because we know that the speed on any network will be different from that of a faux ∞-cloned network. This also poses problems whenever we want

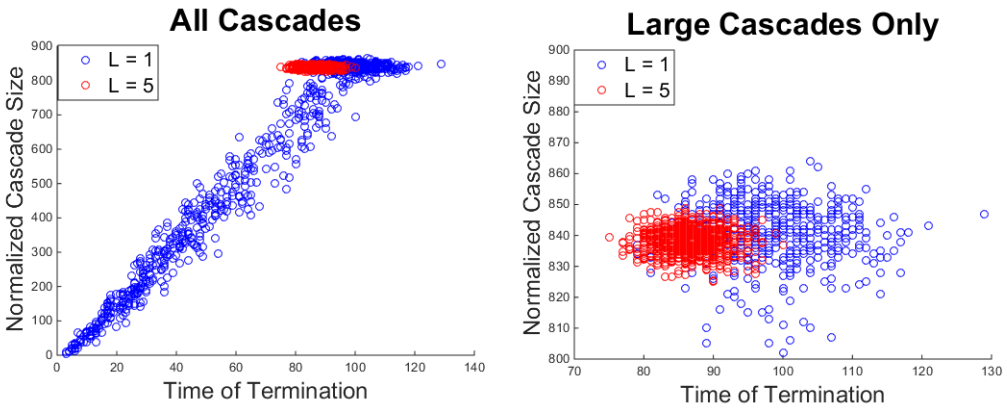

Figure 6.5: We run 1,000 simulations on our toy network (blue) and on a faux 5-cloned version of the network (red). We plot the normalized final cascade size $\frac{\rho N}{L}$ with respect to the termination time. The faux five-cloned graph takes less time to propagate a given distance than the uncloned graph does. Because cascades on the faux 5-cloned network only terminate due to exhaustion, the relevant ranges of cascade sizes and termination times are different for the cloned and uncloned cases. The left panel shows the full range of termination times and cascade sizes while the right panel shows those times and cascade sizes relevant for the cloned network.

to compare statistics relating to the shape of $\rho(x, \tau)$ (such as the standard deviation of the geographic locations of agents that activated at a given time) as these parameters may also differ from a given network to its faux ∞-cloned version. Section 6.3 describes an approach to estimate the transition probabilities from $y_{\tau-1}$ to y_τ for any positive integers $y_{\tau-1}$ and y_τ. Knowledge of the full probability transition matrix will allow us to account for the lower propagation speed and will give us an estimate for the termination probability.

6.3 New Activations as a Markov Chain

We now return our focus to the original non-cloned network. On this network, we want to know the likelihood that some number Y_τ agents activated at time τ conditioned on some number $Y_{\tau-1}$ agents activating at time $\tau - 1$.

In order to calculate these transition probabilities, we run the mean field simulations for enough time that the evolution of the cascade becomes a propagating wave. We can evaluate whether such a steady-state has been reached by comparing statistics about $\rho(x, \tau) - \rho(x, \tau - 1)$ and $\rho(x, \tau - 1) - \rho(x, \tau - 2)$. We want to make sure that the propagation speed and the wave shape are sufficiently similar between those two consecutive time steps. One way to

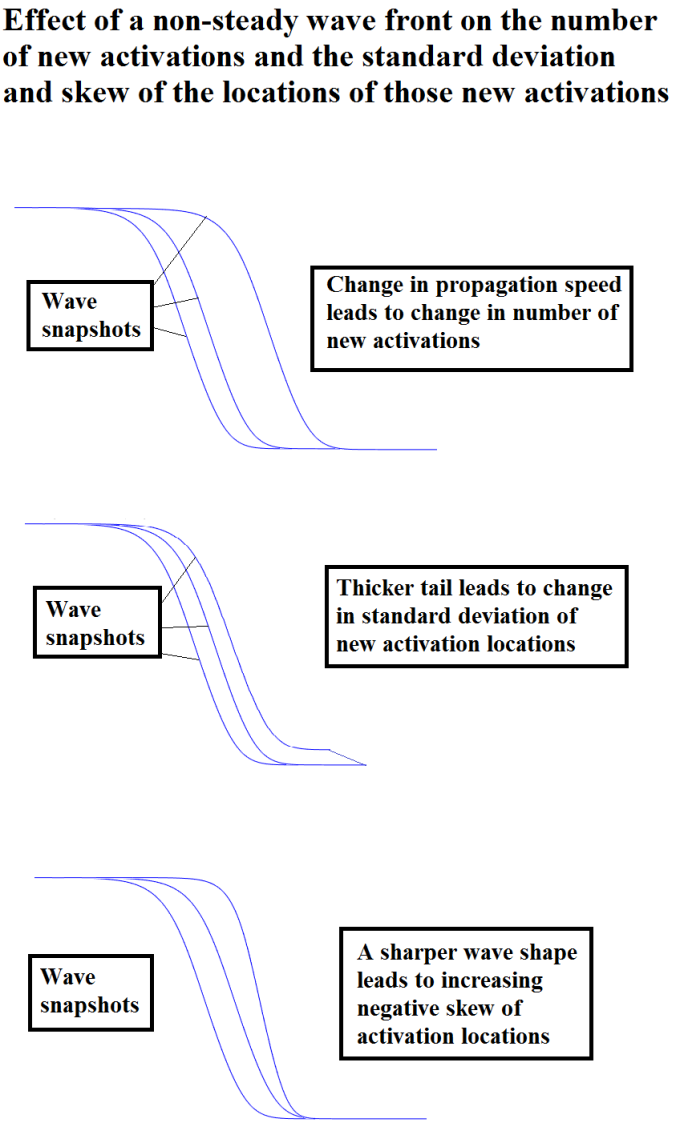

Figure 6.6: Visual illustrations of the relevance of the number of new activations, the standard deviation of their locations, and the skew of their locations when assessing whether a steady wave propagation has been reached.

do this would be to compare $\int_0^w \rho(x,\tau) - \rho(x,\tau-1)dx$ and $\int_0^w \rho(x,\tau-1) - \rho(x,\tau-2)dx$ to compare propagation speeds. Unfortunately, those integrals give us little information about the extent to which the wave shape may be evolving over time. One possible solution is to compare the standard deviation and skew of the probability distribution of the locations of the agents that activated at a given time. Figure 6.6 illustrates why these parameters may be relevant. If the wave shape is changing by widening the region of new activity, the standard deviation will increase, as illustrated in the middle panel, and increasingly steep wave fronts lead to decreasing skewness of the new activation locations. For simplicity, we declare the function $\tilde{\rho}(x,\tau)$ to be $\frac{\rho(x,\tau)-\rho(x,\tau-1)}{\int_0^w \rho(x,\tau)-\rho(x,\tau-1)dx}$. That is, $\tilde{\rho}$ is a normalized distribution of new activations. The standard deviation of the locations of nodes that activated at time τ is then

$$\sigma(\tilde{\rho}) = \sqrt{\int_0^w x^2\tilde{\rho}(x,\tau)dx - (\int_0^w x\tilde{\rho}(x,\tau)dx)^2} \tag{6.10}$$

and the skew is

$$\text{skew}(\tilde{\rho}) = \frac{\int_0^w (x - \int_0^w \hat{x} \times \tilde{\rho}(\hat{x},\tau)d\hat{x})^3 \times \tilde{\rho}(x,\tau)dx}{(\int_0^w x^2 \times \tilde{\rho}(x,\tau)dx - (\int_0^w x \times \tilde{\rho}(x,\tau)dx)^2)^{1.5}} \tag{6.11}$$

Once the number of new activations and the standard deviation and skew of the activation locations at times τ and $\tau+1$ are within some tolerance of each other, we can assume that a steady state has been reached by time τ.

Assume that this simulation took τ time steps. We keep track of the values of $\rho(x,\tau)$ and $T_\tau(k,m,t,x)$ after τ and $\tau-1$ time steps. Because the cascade propagates more quickly on the faux ∞-cloned network than on the original network, there will be some systematic errors in the mean field estimates of these values. $\rho(x,\tau)$ will be lower than its estimated value, and the region where $\rho(x,\tau)$ and $T_\tau(k,m,t,x)$ change the most from $\rho(x,\tau-1)$ and $T_{\tau-1}(k,m,t,x)$ will be further to the left than predicted because the cascade is propagating more slowly from left to right. We are only concerned with the number of agents that activate at a given time, not which ones they are, so we ignore the latter inaccuracy. Our approach can predict the average number of agents that activate at time $\tau-1$ with the formula

$$E[Y_{\tau-1}] = \int_0^w \rho(x,\tau-1) - \rho(x,\tau-2)dx \tag{6.12}$$

where w is the width of the network. This integral is just the difference between the average total number of active agents at time τ and the average number of active agents at time $\tau - 1$. If we know the actual value of $Y_{\tau-1}$, we would like to estimate not only the average value of Y_τ but full probability distribution of the random variable Y_τ. Using the assumption that different inactive agents will activate independently and using a limiting assumption of there being many agents each of which only has a small chance to activate, we can approximate the distribution of Y_τ with a Poisson distribution. We validate this assumption in Figure 6.7. We run 30,000 simulated cascades on the uncloned version of our toy network. For each of several values of $Y_{\tau-1}$, we compare the empirical distribution of Y_τ to the Poisson distribution with the same mean. The solid curves are the empirical distributions and the dotted curves are the Poisson distributions. The agreement between the empirical and theoretical curves is excellent.

If we were to run the faux ∞-cloned simulation for an additional time step, we would want to know the likelihood of a neighbor of an inactive agent at point x activating at time $\tau - 1$, making the agent at point x more likely to activate at time τ. We weight all nodes equally, regardless of the degree of each node. We use equation (6.6) to find that the probability that an agent that was inactive at $\tau - 2$ and within range $R(x)$ of x activated at $\tau - 1$ is

$$\bar{q}(x, \tau - 1) = \frac{\int_{R(x)} \rho(\hat{x}, \tau - 1) - \rho(\hat{x}, \tau - 2) d\hat{x}}{\int_{R(x)} (1 - \rho(\hat{x}, \tau - 2)) d\hat{x}} \tag{6.13}$$

This equation uses the assumption that all agents are weighted equally, regardless of their degrees. This formula only gives the *expected* fraction of inactive agents within range of a given agent that will activate on the subsequent time step without any further information. If we know $Y_{\tau-1}$ to take a value smaller or larger than its expected value $E[Y_{\tau-1}]$, we should expect $\bar{q}(x, \tau - 1)$ to be affected accordingly. (If there were more activations overall at time $\tau - 1$, this would increase the likelihood of a specific neighbor of a specific agent activating at time $\tau - 1$.) We would also expect $\rho(x, \tau - 1)$ to increase and $\rho(x, \tau - 2)$ to slightly decrease with increasing $y_{\tau-1}$. (If more agents activate at time $\tau - 1$ exactly, we would expect more agents to activate *at or before* time $\tau - 1$, and slightly more agents to be available for activation at time $\tau - 1$.) With this consideration, we condition $\bar{q}(x, \tau - 1)$, $\rho(x, \tau - 1)$, and $\rho(x, \tau - 2)$ on $y_{\tau-1}$. There is a distinction between the probability of an agent at location x

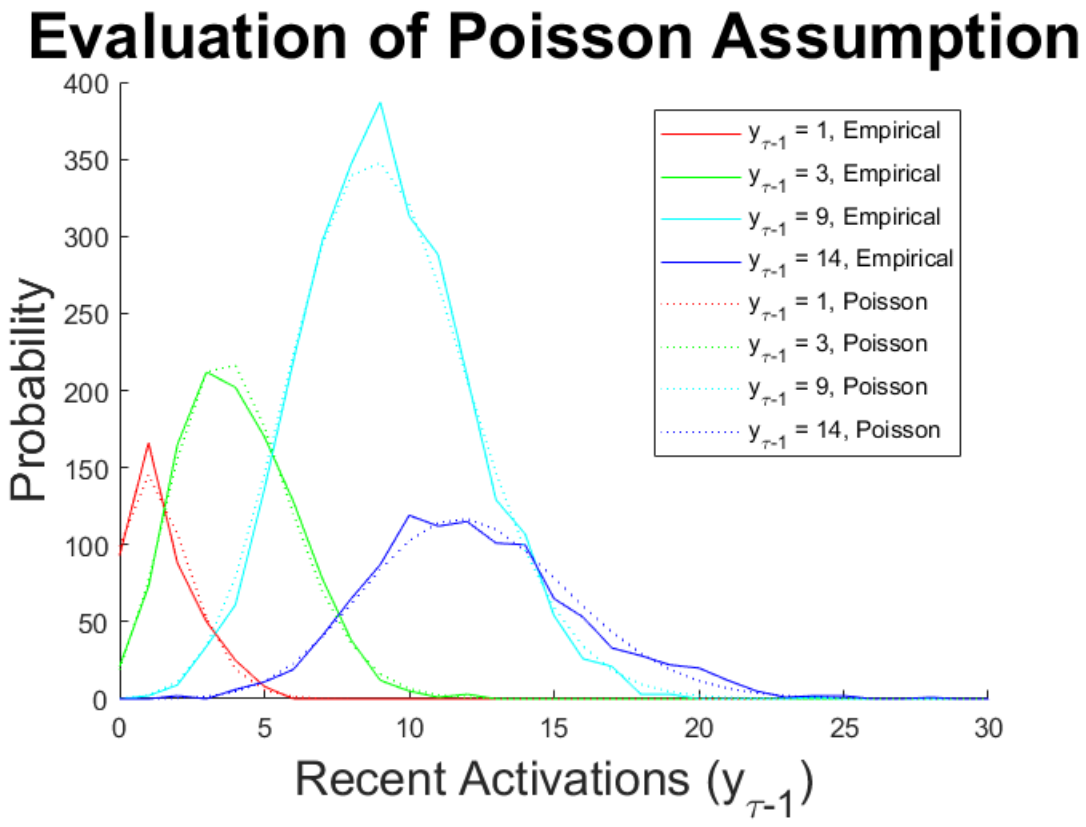

Figure 6.7: When the number of recent activations $y_{\tau-1}$ is $1, 3, 9, 14$ we compare the empirical distribution of immediately upcoming activations Y_τ gathered over 30,000 simulated cascades to the Poisson distribution with the same mean. The solid curves are the empirical distributions and the dotted curves are the Poisson distributions.

being active at time $\tau - 1$ with no knowledge of the number of agents $y_{\tau-1}$ that activated at time $\tau - 1$ exactly and the probability of an agent at location x being active at time $\tau - 1$ conditioned on a value of $y_{\tau-1}$. To distinguish between these two functions, we denote the former by $\rho(x, \tau - 1)$ and the latter by $\rho_{y_{\tau-1}}(x, \tau - 1)$.

To determine how knowledge of $y_{\tau-1}$ affects the probability of an agent at location x being active at each of times $\tau - 1$ and $\tau - 2$, we assume that the agents at each location x behave independently, and have probability $\rho(x, \tau - 2)$ to be active at time $\tau - 2$. We assume that those agents that are inactive at time $\tau - 2$ have probability $\frac{\rho(x,\tau-1)-\rho(x,\tau-2)}{1-\rho(x,\tau-2)}$ to activate at time $\tau - 2$ exactly. While this assumption provides enough information to determine the conditional probabilities $\rho_{y_{\tau-1}}(x, \tau - 2)$ and $\rho_{y_{\tau-1}}(x, \tau - 1)$ exactly, the resulting formula is too computationally intensive to use. To resolve this issue, we approximate $\rho_{y_{\tau-1}}(x, \tau - 2)$

and $\rho_{y_{\tau-1}}(x, \tau-1)$ by assuming that the number of agents not at location x that activate at any given time is Poisson distributed. It is easiest to consider the effects of $y_{\tau-1}$ on the local probability of recent activation $\rho(x, \tau-1) - \rho(x, \tau-2)$ and the local probability of being already active $\rho(x, \tau-2)$ separately. At this point, it is convenient to define the functions $\rho_{\text{new}}(x, \tau-1)$ and $\rho_{\text{new},y_{\tau-1}}(x, \tau-1)$ respectively as the probability of an agent at x activating at time $\tau-1$ exactly and the conditional probability of an agent at x activating at time $\tau-1$ exactly given that $y_{\tau-1}$ total agents activated at time $\tau-1$. As mentioned, without knowledge of $y_{\tau-1}$, we consider the behaviors of each of the N agents to be independent. Thus, the probabilities of each agent activating at time $\tau-1$ exactly are independent Bernoulli random variables, each with success probability approximately $\rho(x, \tau-1)-\rho(x, \tau-2)$, where x is the location of the agent. If we know that the number of agents that activated at time $\tau-1$ is some $y_{\tau-1}$ then this introduces some dependence between the behaviors of different agents. Unfortunately, there is no tractable way to calculate the probability under such circumstances that a specific agent at some x was one of the $y_{\tau-1}$ agents that activated at time $\tau-1$ exactly. (The only option is to calculate the probability of each possible permutation of $y_{\tau-1}$ of the N total agents activating and compare the combined probabilities of those permutations including the agent at x to the combined probabilities of those permutations which do not include that agent.) Unable to calculate the exact result, we approximate the number of agents *besides* the one at x that activate at time $\tau-1$ with a Poisson distribution with the same mean. (Figure 6.7 validates the assumption that the number of recent activations is approximately Poisson.) Under this assumption, the probability $P_{\text{Act},y_{\tau-1}}(x, \tau-1)$ of a given agent at x activating at time $\tau-1$ and there being $y_{\tau-1}$ activations overall at time $\tau-1$ is

$$P_{\text{Act},y_{\tau-1}}(x, \tau-1) = \rho_{\text{new}}(x, \tau-1) \times (PO(y_{\tau-1}-1, E[Y_{\tau-1}] - \rho_{\text{new}}(x, \tau-1)) \quad (6.14)$$

where $PO(x, \lambda)$ is the probability of a Poisson random variable with mean λ having value x. Similarly,the probability $P_{\text{NotAct},y_{\tau-1}}(x, \tau-1)$ of the agent at x not activating at time $\tau-1$ exactly and there being $y_{\tau-1}$ agents that activated at time $\tau-1$ is

$$P_{\text{NotAct},y_{\tau-1}}(x, \tau-1) = (1-\rho_{\text{new}}(x, \tau-1)) \times (PO(y_{\tau-1}, E[Y_{\tau-1}] - \rho_{\text{new}}(x, \tau-1))). \quad (6.15)$$

Comparing the likelihood of an agent at x activating at time $\tau - 1$ and there being $y_{\tau-1}$ total activations at time $\tau - 1$ to the likelihood of an agent at x not activating at time $\tau - 1$ and there being $y_{\tau-1}$ agents that activated at time $\tau - 1$ across the whole network, we get

$$\rho_{\text{new},y_{\tau-1}}(x, \tau - 1) = \frac{P_{\text{Act},y_{\tau-1}}(x, \tau - 1)}{P_{\text{Act},y_{\tau-1}}(x, \tau - 1) + P_{\text{NotAct},y_{\tau-1}}(x, \tau - 1)}. \tag{6.16}$$

Unfortunately, the Poisson approximation is not perfect, especially for particularly high or particularly low values of $y_{\tau-1}$. This could lead to the paradoxical result

$$\int_0^w \rho_{\text{new},y_{\tau-1}}(x, \tau - 1)dx \neq y_{\tau-1}. \tag{6.17}$$

Because the inaccuracy lies in the Poisson terms, which are nearly identical for all x, we assume that the *relative* likelihoods of agents at different locations activating at time τ are accurately determined by (6.16), so we normalize $\rho_{\text{new},y_{\tau-1}}(x, \tau - 1)$ by a factor of $\frac{y_{\tau-1}}{\int_0^w \rho_{\text{new},y_{\tau-1}}(x,\tau-1)dx}$. to get

$$\bar{\rho}_{\text{new},y_{\tau-1}}(x, \tau - 1) = \frac{\rho_{\text{new},y_{\tau-1}}(x, \tau - 1) \times y_{\tau-1}}{\int_0^w \rho_{\text{new},y_{\tau-1}}(x, \tau - 1)dx}. \tag{6.18}$$

For some extremely high values of $y_{\tau-1}$ this could cause $\bar{\rho}_{\text{new},y_{\tau-1}}(x, \tau - 1) > 1$. In those cases, we would cap $\bar{\rho}_{\text{new},y_{\tau-1}}(x, \tau - 1)$ at 1. However, these values of $y_{\tau-1}$ are so unrealistically high (around 60 on our toy network) that they are astronomically unlikely. In our 30,000 simulated cascades y_{20} never exceeded 29, so this is of minimal practical concern. As will be explained later in this section, it will be necessary to set some maximum realistic number $y_{\max}$ of activations per unit time for our method to be computationally tractable. Because of the discrepancy between the conditional $\bar{\rho}_{\text{new},y_{\tau-1}}(x, \tau - 1)$ and the unconditional $\bar{\rho}_{\text{new}}(x, \tau - 1)$ there should be some discrepancy between $\rho_{y_{\tau-1}}(x, \tau - 2)$ and $\rho(x, \tau - 2)$. If an agent activated at time $\tau - 1$ exactly then it must have been inactive at time $\tau - 2$ and active at time $\tau - 1$. We calculate that this happens with probability $\bar{\rho}_{\text{new},y_{\tau-1}}(x, \tau - 1)$. If an agent did not activate at time $\tau - 1$ exactly then either it was already active at time $\tau - 2$ or it was inactive at time $\tau - 1$. In such a case, because we presume the agents to behave independently, the fact that some specific number $y_{\tau-1}$ of other agents activated at time $\tau - 1$ exactly does not affect the probabilities of an agent at location x having activated

by time $\tau - 2$ or still being inactive at time $\tau - 1$ given that the agent did not activate at time $\tau - 1$ exactly. Put another way, $1 - \bar{\rho}_{\text{new}}(x, \tau - 1)$ represents that probability that an agent at location x did not activate at time $\tau - 1$ exactly and $1 - \rho(x, \tau - 2)$ represents the probability that the agent did not activate at time $\tau - 1$ exactly *and* did not activate before then. This means that the probability that an agent activated before time $\tau - 1$ given that it did not activate at time $\tau - 1$ exactly is $\frac{\rho(x,\tau-2)}{1-\bar{\rho}_{\text{new}}(x,\tau-1)}$. We presume that the specific value $y_{\tau-1}$ does not affect the probability of this last condition. That is, if we know that an agent at location x did not activate at some precise time $\tau - 1$ then we presume that knowledge of $y_{\tau-1}$ does not change the probability that the agent activated earlier. Because knowledge of $y_{\tau-1}$ changes the probability of an agent not activating at time $\tau - 1$ exactly from $1 - \bar{\rho}_{\text{new}}(x, \tau - 1)$ to $1 - \bar{\rho}_{\text{new},y_{\tau-1}}(x, \tau - 1)$ and the conditional probability that an agent that did not activate at time $\tau - 1$ activated before time $\tau - 1$ remains the same at $\frac{\rho(x,\tau-2)}{1-\bar{\rho}_{\text{new}}(x,\tau-1)}$, we get the formula

$$\rho_{y_{\tau-1}}(x, \tau - 2) = \frac{\rho(x, \tau - 2)}{1 - \bar{\rho}_{\text{new}}(x, \tau - 1)} \times (1 - \bar{\rho}_{\text{new},y_{\tau-1}}(x, \tau - 1)). \tag{6.19}$$

Because the probability of an agent at location x being active at time $\tau - 1$ is just the sum of the probability of its being active at time $\tau - 2$ and the probability of its activating at time $\tau - 1$ exactly, we get

$$\rho_{y_{\tau-1}}(x, \tau - 1) = \rho_{y_{\tau-1}}(x, \tau - 2) + \rho_{\text{new},y_{\tau-1}}(x, \tau - 1). \tag{6.20}$$

Now consider the probability $T_{\tau-1}(k, m, t, x)$ that an agent at location x was inactive at time $\tau - 1$ and had degree k, m previously active neighbors, and threshold t. This probability also changes when $y_{\tau-1}$ is determined. For this reason, we distinguish between the unconditional $T_{\tau-1}(k, m, t, x)$ and the conditional probability $T_{\tau-1,y_{\tau-1}}(k, m, t, x)$ of an agent at location x being inactive at time $\tau - 1$ and having degree k, m previously active neighbors and threshold t given that $y_{\tau-1}$ agents activated at time $\tau - 1$. To determine the relationship between the conditional $T_{\tau-1,y_{\tau-1}}(k, m, t, x)$ and the unconditional $T_{\tau-1}(k, m, t, x)$, we use an assumption similar to that used to develop (6.19). If we know that the agent at location x was inactive at time $\tau - 1$, we presume that the fact that exactly $y_{\tau-1}$ other agents activated at time $\tau - 1$ does not change the probability that the agent at x has any given degree, number of already active neighbors, or threshold given that it was inactive at time $\tau - 1$.

This gives us

$$T_{\tau-1,y_{\tau-1}}(k,m,t,x) = \frac{T_{\tau-1}(k,m,t,x)}{1-\rho(x,\tau-1)} \times (1-\rho_{y_{\tau-1}}(x,\tau-1)) \tag{6.21}$$

Now consider the probability $p(n|k,m,x,\tau,y_{\tau-1})$, the probability that an inactive agent at location x with k neighbors, m of which were already active by time $\tau-2$ received n new spikes at time $\tau-1$. To figure out the relationship between $p(n|k,m,x,\tau,y_{\tau-1})$ and $y_{\tau-1}$, we address the question of how the unconditional $p(n|k,m,x,\tau)$ (without knowledge of $y_{\tau-1}$ is related to $p(n|k,m,x,\tau,y_{\tau-1})$ for an arbitrary $y_{\tau-1}$. As indicated in Figure 6.8, some edges are more likely to transmit spikes than others. Consider edge a, near the left end of the map. At time $\tau-1$ it is most likely that the agents connected by that edge are both already active, so one would not activate again and send a spike to the other. At the other extreme, consider edge c near the right end of the map. The wave has hardly affected this part of the network by time $\tau-1$, so it is unlikely that one of these agents will activate and send a spike to the other. Looking at the evolution of the cascade between $\tau-2$ and $\tau-1$ in the vicinity of edge b, we notice there is a much greater likelihood of one of its two nodes activating at time $\tau-1$ exactly and sending a spike to the other. Suppose we knew the probability of each edge transmitting a spike, and thus, the expected value of $Y_{\tau-1}$. Suppose we were given the actual number $y_{\tau-1}$ of agents that activated across the entire network. This number $y_{\tau-1}$ would affect the likelihood of any given edge transmitting a spike. A higher value of $y_{\tau-1}$ should increase the probability of each edge transmitting a spike, and a lower value of $y_{\tau-1}$ should decrease it.

Suppose that there were $y_{\tau-1}$ agents that activated at time $\tau-1$ across the entire network and that an inactive agent at position x had degree k and m agents that were already active at time $\tau-1$. We would like to estimate the probability that exactly n of the agent's neighbors were among the $y_{\tau-1}$ agents that activated at time $\tau-1$. To do this rigorously and precisely, we would need to consider the possibility that each of the N agents was one of the $y_{\tau-1}$ total agents that activated in the most recent time step, then consider the likelihood that exactly n of them were adjacent to the agent at x. This is too computationally intensive, so we make a simplifying assumption. We assume that the inactive neighbors of the agent at x have some common probability $\omega_1(x,\tau-1)$ to be among the most recently activated agents and that the other agents have some common probability $\omega_2(x,\tau-1)$ to activate at time τ exactly. We can determine the probability distribution of

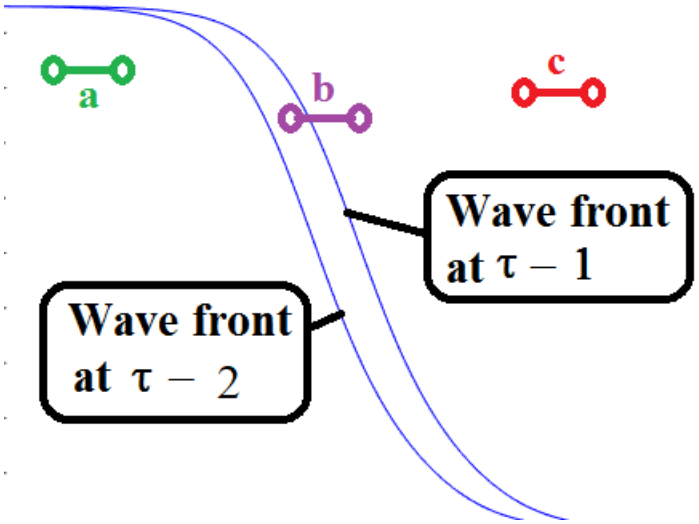

Figure 6.8: An illustration of why some edges are more likely to transmit spikes than other. Edge b (in purple) is more likely to transmit a spike at time $\tau - 1$ than edge a (in green) or edge c (in red).

n, the number of recently activated neighbors of the agent at x, through Fisher's noncentral hypergeometric distribution.

The neighbors of the agent at x must all be within range of x. Weighting all agents within range equally, we find that the likelihood of a given one of them activating at time $\tau - 1$ conditioned on its being inactive at time $\tau - 2$ is

$$\omega_1(x, \tau - 1) = \frac{\int_{R(x)} \bar{\rho}_{\text{new},y_{\tau-1}}(\hat{x}, \tau - 1)d\hat{x}}{\int_{R(x)}(1 - \rho_{y_{\tau-1}}(\hat{x}, \tau - 2))d\hat{x}}. \tag{6.22}$$

Of the N agents in the network, almost exactly $2r - k - 1$ of them are inside the range of the agent at x, but not neighbors of the agent at x and not the agent at x itself. Due to the randomness in the locations of the agents, this number could range from $2r - k - 2$ to $2r - k$. Because this variation is slight, we approximate this number by its modal value, $2r - k - 1$. Each of them has probability $\omega_{2,in}(x, \tau - 1)$ to be recently activated, calculated

$$\omega_{2,in}(x, \tau - 1) = \frac{\int_{R(x)} \bar{\rho}_{\text{new},y_{\tau-1}}(\hat{x}, \tau - 1)d\hat{x}}{\int_{R(x)} 1 d\hat{x}}. \tag{6.23}$$

Note the difference in denominators between (6.22) and (6.23). This difference arises from the fact that in (6.22) the $k-m$ previously inactive neighbors of the agent at x are known

to be previously inactive, and thus cannot be among those agents that did not activate at τ on account of prior activation. In contrast, in (6.23), only recently activated agents within range count toward the numerator, but *all* agents count towards the denominator, whether they are already active, recently activated, or still inactive. In contrast, some of the non-neighbors of the agent at x within range of x could have been previously active, and thus have a lower likelihood of activating at time τ.

In addition to the almost exactly $2r - k - 1$ non-neighbor agents within range of the agent at x, there are almost exactly $N - 2r$ agents outside the range of x. (This number can range from $N - 2r - 1$ to $N - 2r + 1$. We approximate this number with its modal value of $N - 2r$.) Each of these will activate at time $\tau - 1$ with probability

$$\omega_{2,out}(x, \tau - 1) = \frac{\int_0^w \bar{\rho}_{\text{new},y_{\tau-1}}(\hat{x}, \tau - 1)d\hat{x} - \int_{R(x)} \bar{\rho}_{\text{new},y_{\tau-1}}(\hat{x}, \tau - 1)d\hat{x}}{\int_0^w 1 d\hat{x} - \int_{R(x)} 1 d\hat{x}}, \tag{6.24}$$

where w is the width of the network. $\omega_2(x, \tau - 1)$ can be calculated by taking the weighted average

$$\omega_2(x, \tau - 1) = \frac{(N - 2r)\omega_{2,out}(x, \tau - 1) + (2r - k - 1)\omega_{2,in}(x, \tau - 1)}{N - k - 1}. \tag{6.25}$$

With the values of k, m, $\omega_1(x, \tau - 1)$, $\omega_2(x, \tau - 1)$, and the number of recently activated agents y, we would like to know the likelihood of exactly n of the $k - m$ previously inactive neighbors of the agent at x activating at time τ. If $\omega_1(x, \tau - 1) = \omega_2(x, \tau - 1)$ we can model this with the well-known hypergeometric distribution. In the more general case where $\omega_1(x, \tau - 1)$ and $\omega_2(x, \tau - 1)$ are allowed to differ, we need to use a generalization of the hypergeometric distribution, called the Fisher noncentral hypergeometric distribution, which gives us the following probability distribution for the number of spikes n received by an agent at time $\tau - 1$.

$$p(n|k, m, x, \tau - 1, y_{\tau-1}) = \frac{\binom{k-m}{n}\binom{N-k-1}{y_{\tau-1}-n}\omega(x, \tau - 1)^n}{\sum_{\hat{n}=0}^{min(k-m, y_{\tau-1})}\binom{k-m}{\hat{n}}\binom{N-k-1}{y_{\tau-1}-\hat{n}}\omega(x, \tau - 1)^{\hat{n}}} \tag{6.26}$$

where $\omega(x, \tau - 1) = \frac{\omega_1(x,\tau-1)}{\omega_2(x,\tau-1)}$. The likelihood $\tilde{\rho}_{\text{new},y_{\tau-1}}(x, \tau)$ of an agent at x activating at time τ given that $y_{\tau-1}$ agents activated at the *previous* time step, at time $\tau - 1$ is then

$$\tilde{\rho}_{\text{new},y_{\tau-1}}(x,\tau) = \sum_k \sum_t \sum_{m=0}^{\min(k,t-1)} \sum_{n=t-m}^{k-m} T_{\tau-1}(k,m,t,x,y_{\tau-1}) \times p(n|k,m,x,\tau,y_{\tau-1}). \quad (6.27)$$

The average total number of agents that activate at time τ is then

$$E[Y_\tau|Y_{\tau-1} = y_{\tau-1}] = \int_0^w \tilde{\rho}_{\text{new},y_{\tau-1}}(x,\tau))dx \quad (6.28)$$

Using the assumption that we can use a Poisson distribution to model the distribution of Y_τ, we get the equation

$$p(Y_\tau = y_\tau|Y_{\tau-1} = y_{\tau-1}) = \frac{\lambda^{y_\tau} \times e^{-\lambda}}{y_\tau!} \quad (6.29)$$

where

$$\lambda = \int_0^w \tilde{\rho}_{\text{new},y_{\tau-1}}(x,\tau)dx. \quad (6.30)$$

This lets us compute the probability transition matrix of the number of activations from time $\tau - 1$ to time τ, as $\mathbf{P}_{y_{\tau-1}+1,y_\tau+1} = p(Y_\tau = y_\tau|Y_{\tau-1} = y_{\tau-1})$. Unfortunately, this would result in an infinite probability transition matrix. To be computationally tractable, our procedure must be restricted to finite matrices. We set some maximum value $y_{\max}$ of $y_{\tau-1}$ and y_τ. For our simulations, we set $y_{\max} = 29$, making the probability transition matrix $\mathbf{P}$ a 30×30 matrix. In our $30,000$ simulated cascades, y_{20} never exceeded 29, suggesting that our maximum degree is a realistic one. To keep $\mathbf{P}$ a stochastic matrix (that is to make sure the rows sum to exactly 1) we set

$$\mathbf{P}_{y_{\tau-1}+1,y_{\max}+1} = 1 - \sum_{j=1}^{y_{\max}} \mathbf{P}_{y_{\tau-1}+1,y_j}. \quad (6.31)$$

That is, if we would assume $y_\tau > y_{\max}$ we instead assume that $y_\tau = y_{\max}$. This has a negligible effect on the final cascade size distribution because the likelihood of $y_\tau > y_{\max}$ is so small.

We now need a set of assumptions on how to approximate the CDF of cascade sizes G from the probability transition matrix $\mathbf{P}$. Suppose that the probability transition matrix $\mathbf{P}$ has quasistationary distribution $\mathbf{q}$, calculated by taking the left eigenvector (corresponding

to the highest positive eigenvalue) of the submatrix of the probability transition matrix P with the rows and columns of all absorbing states removed, as described in [5]. We then presume the system to be at the quasistationary distribution at time $\tau - 1$ and progress through one more time step according to the probability transition matrix $\mathbf{P}$ to get the vector $\hat{\mathbf{q}} = \mathbf{q}\ \mathbf{P}$. We can now calculate the termination probability which we called C_2 in Chapter 5, but now makes more sense to call p_0. The resulting equation is

$$p_0 = \begin{bmatrix} 1 & 0 & 0 & \dots & 0 \end{bmatrix} \hat{\mathbf{q}}. \tag{6.32}$$

We also want to know the mean propagation speed of the cascade $E[Y]$, which we presume to be the mean number of new activations per unit time according to the quasistationary distribution. This gets us the formula

$$E[Y] = \begin{bmatrix} 0 & 1 & 2 & \dots & y_{\max} \end{bmatrix} \mathbf{q}, \tag{6.33}$$

where $y_{\max}$ is the value that we take as the maximum realistic number of activations at any given time. Assuming that there are N_0 initial seeds, we get the following formula for $G_{\text{spon}}(z)$, the probability that the cascade will terminate before reaching size z

$$G_{\text{spon}}(z) = \begin{cases} 0 & \text{for } z < N_0 \\ 1 - (1 - p_0)^{\frac{z - N_0}{E[Y]}} & \text{for } z \geq N_0 \end{cases}. \tag{6.34}$$

There is the possibility that the cascade will terminate through exhausting the supply of available agents, causing a cascade of the approximate size predicted by Gleeson and Cahalane in [12]. In such a case, we presume the number of active agents to follow an approximate Gaussian distribution. The N_0 seeds are known to be active and the $N - N_0$ non-seeds are presumed to be independently active with probability ρ_{sat}, the saturation density, which can be calculated with (3.34). This gets us

$$G_{\text{exst}}(z) \approx \begin{cases} 0 & \text{for } z < N_0 \\ \Phi(z, N_0 + (N - N_0) \times \rho_{\text{sat}}, (N - N_0) \times \rho_{\text{sat}} \times (1 - \rho_{\text{sat}})) & \text{for } z \geq N_0 \end{cases}, \tag{6.35}$$

where $\Phi(x, \mu, \sigma^2)$ represents the CDF of a Gaussian distribution with mean μ and

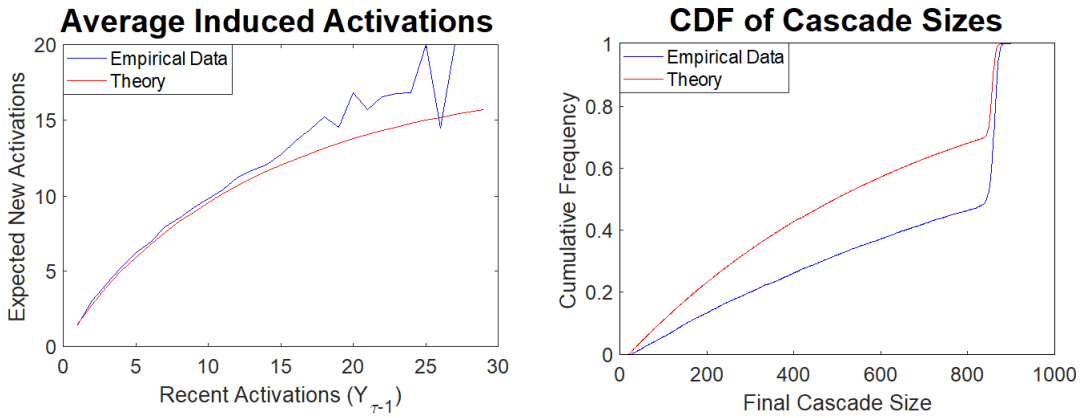

Figure 6.9: Comparisons of the $Y_{\tau-1}$-to-$E[Y_\tau]$ curve (left) and CDF of final cascade size (right) predicted by our three-part method to their empirical values. The empirical data come from 10,000 simulated cascades using our toy network.

variance σ^2 as a function of x. For a cascade to survive to size z it must not terminate before size z either by spontaneous termination or through saturation. We presume these two possible reasons for termination to be independent, giving us

$$G(z) = 1 - (1 - G_{\text{spon}}(z)) \times (1 - G_{\text{exst}}(z)). \tag{6.36}$$

We use these assumptions to approximate the CDF of cascade sizes, G. The left panel of Figure 6.9 compares this approximation to the empirical result found from 10,000 simulated cascades. Our method predicts a far greater likelihood of early termination than the simulations suggest. We discuss methods to remedy this problem in Chapter 7.

CHAPTER 7
NECESSARY MODIFICATIONS FOR THE METHOD TO BE ACCURATE

The results of the coarse approach of Section 6.3 are plotted in Figure 6.9 which compares the predicted final cascade size to the actual cascade size on the same toy network we have been using throughout this work. The prediction is wildly inaccurate. We already know that if the transition probabilities are accurate, the cascade size distribution prediction will be accurate, as shown in Figure 5.7 above. Additionally, Figure 6.7 shows us that if we can accurately determine $E[Y_\tau|Y_{\tau-1} = y_{\tau-1}]$ then we can approximate each individual $P(Y_\tau = y_\tau|Y_{\tau-1} = y_{\tau-1})$. That is, if we can accurately determine the average number of new activations given the number of new activations one time step prior, we can determine the probability transition matrix, and from there, use this matrix to estimate the final cascade size distribution.

Because the final prediction is inaccurate, we can infer that there is some inaccuracy in our estimates of $E[Y_\tau|Y_{\tau-1} = y_{\tau-1}]$. Otherwise, we should have an accurate estimate of the probability transition matrix which would lead to an accurate estimate of the CDF of final cascade sizes. Section 7.1 describes a possible source of error, while sections 7.2, 7.3, and 7.4 develop a remedy to that error.

7.1 The Need to Account for Total Spikes Sent

An agent with k total neighbors will activate when its number of active neighbors meets or exceeds its threshold t. This means that if it activates the number of spikes it sends out to inactive agents is at most $k - t$, but not necessarily $k - t$ exactly. First, the agent could receive more spikes than it needs to activate. Second, it may activate at the same time as one or more of its neighbors. In this case, while those neighbors would not count toward its threshold, the spikes it sends to them will not affect the cascade propagation in any way, as the receiving agents will be active by the time they receive those spikes. We refer to the number of spikes sent to agents that were not already active as *relevant spikes*. Figure 7.1 illustrates the distinctions between edges that transmit relevant spikes and edges that

transmit each of several types of irrelevant spikes. The black nodes with the yellow numbers in them represent agents which activated at time $\tau - 1$. In Figure 7.1 there are two such nodes, labeled a and b. The yellow numbers represent the response thresholds of those nodes. In Figure 7.1, both of these numbers are 2. Black nodes that do not have yellow numbers in them represent agents that activated before time $\tau - 1$ and white nodes represent agents that had not activated by time $\tau - 1$. Node a is adjacent to two nodes that were already active before time $\tau - 1$, three nodes that had not activated by time $\tau - 1$, and to node b, which activated at time $\tau - 1$. While node a has six neighbors, it can only influence three of them to activate, as the other three activated before it could send them each a spike. For this reason, node a only sends out three relevant spikes when it activates. Similarly, node b is adjacent to seven other nodes, but can only influence three of them to activate. Since both nodes a and b had response threshold 2, they need at least two neighbors to activate at least one time step prior to their own activations. In this case, node a is adjacent to exactly two such nodes and node b is adjacent to three, one more than is necessary. (We can infer from this that node b received at least two spikes between $\tau - 2$ and $\tau - 1$ or it would have activated before time $\tau - 1$, not at time $\tau - 1$ exactly.) Because node b received enough spikes all at once to push it over its threshold (as opposed to meeting its threshold exactly) we say that it received an *overshot spike* in addition to the two spikes that were required for its activation. (There is ambiguity in how to choose which two of the three spikes sent to node b before it activated to classify as "required for activation" and which one to classify as an "overshot spike." This distinction is not relevant for this calculation.) Furthermore, because nodes a and b are adjacent, each one serves as a neighbor of the other which could not be influenced to activate. That is, node a cannot influence node b to activate because by the time node a sends a spike to node b, node b is already active and vice versa. Such a pair of agents is referred to as a pair of *simultaneous activations*.

Obviously, if more spikes are sent to inactive agents across the network, these spikes are likely to combine to induce more agents to activate on the following time step. For this reason, we are very interested in the number of relevant spikes sent out across the network at time τ. Suppose we know that some number $y_{\tau-1}$ agents activated at time $\tau - 1$. We use the function $U(y_{\tau-1})$ to denote the average number of relevant spikes sent by each of them. We can determine $U(y_{\tau-1})$ if we know $\mu_{\text{mod}}(y_{\tau-1})$, the average degree of agents that activated at time $\tau - 1$, and subtract the mean threshold $t_{\text{mod}}(y_{\tau-1})$ of those agents

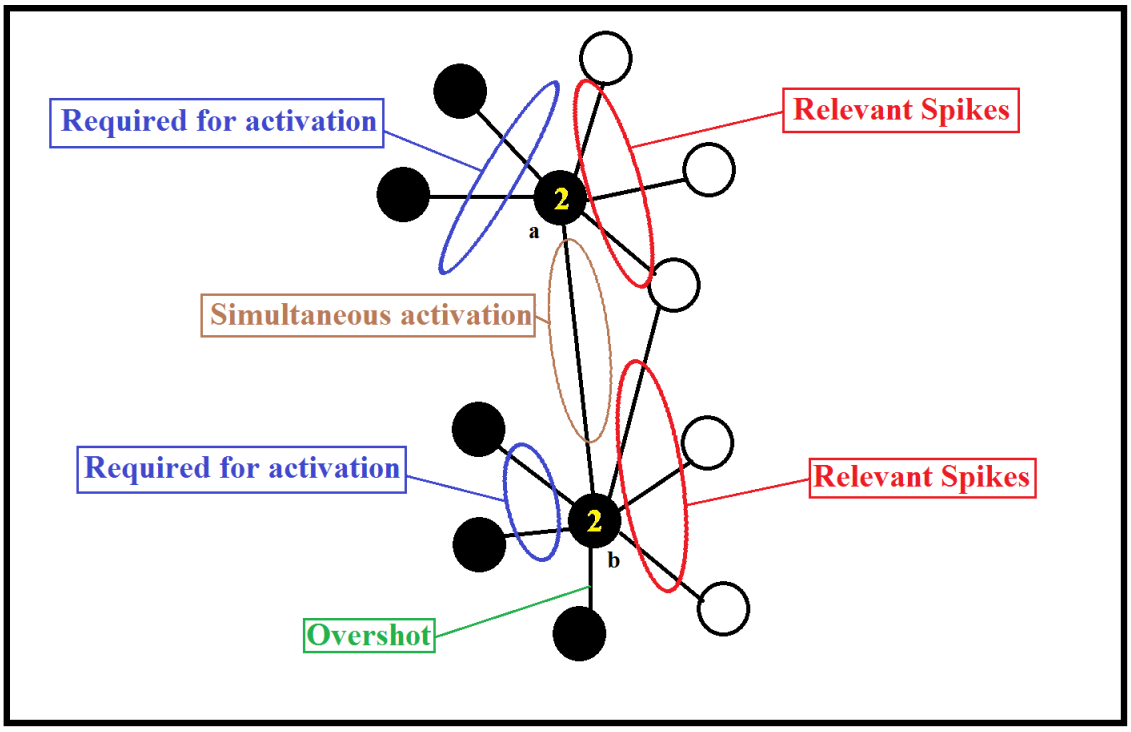

Figure 7.1: A visual representation of the distinction between relevant and irrelevant spikes.

(because this means that on average t_{mod} of those neighbors must have been active before time $\tau - 1$), the expected number of overshot spikes per agent $V(y_{\tau-1})$ (which would have originated from agents that activated at time $\tau - 2$), and the expected number $W(y_{\tau-1})$ of spikes sent from agents which activated at time $\tau - 1$. For a given $y_{\tau-1}$, we can estimate the corresponding value of $U(y_{\tau-1})$ by estimating the average degree of the recently activated agents and subtracting the expected number of irrelevant spikes. The resulting formula is

$$U(y_{\tau-1}) = \mu_{\text{mod}}(y_{\tau-1}) - t_{\text{mod}}(y_{\tau-1}) - V(y_{\tau-1}) - W(y_{\tau-1}) \tag{7.1}$$

We now turn our focus to estimating each of the four quantities on the right side of (7.1). Because we are only concerned with agents that activate after the wave front stabilizes, we do not expect μ_{mod} to depend on τ. Additionally, because agents of higher degree are more likely to activate than agents of lower degree, we expect $\mu_{\text{mod}} > \mu$.

To estimate μ_{mod}, we presume that any agents whose response thresholds do not exceed their degrees will activate eventually. (On our toy network, this is shown to be approximately

true, as the expected fraction of agents that will activate in a nearly-complete cascade is 0.951 and the expected fraction of agents whose response thresholds do not exceed their degrees is 0.959.) We can then take the expectation of the degree of the agents whose response thresholds do not exceed their degrees. Assuming that the degree distribution of agents on the network is Poisson, we get the formula

$$\mu_{\text{mod}} = \frac{\sum_k k \times PO(k,\mu)F(k)}{\sum_k PO(k,\mu)F(k)} \tag{7.2}$$

where $PO(k,\mu)$ is the probability that a Poisson random variable with mean μ takes the value k. $PO(k,\mu)$ can be calculated with the formula

$$PO(k,\mu) = \frac{e^{-\mu}\mu^k}{k!}. \tag{7.3}$$

Note that, according to our approximation, μ_{mod} is constant across all y_τ. Figure 7.2 compares the theoretical value of μ_{mod} to the empirically approximated average across 10,000 simulated cascades. The approximation is reasonably accurate. While there is some systematic error, the discrepancy is very slight.

According to this assumption, an agent will transmit, on average, μ_{mod} spikes if it activates. Some of those spikes will be irrelevant. If the agent had threshold t, then we know that it had to send at least t irrelevant spikes, as it needed at least t active neighbors before it could activate. To estimate the expected number of spikes that were irrelevant for this reason, we would want to know the average threshold of agents that activated at a given time $\tau - 1$. Because we assume that the wave front has already stabilized, we assume this conditional average threshold is independent of τ. To approximate t_{mod}, the average response threshold of all agents that activated at a given point in time we use the same assumption that all agents with enough total neighbors to activate will eventually activate. This means that we would find the average response threshold of those agents whose response thresholds are sufficiently low that they would ever be able to activate. As such, we calculate the expected response threshold T of those agents whose response thresholds do not exceed their degrees K, $E[T|T \leq K]$. Our formula is

$$t_{\text{mod}} \approx E[T|T \leq K] = \frac{\sum_t t \times (1 - \sum_{\hat{k}=0}^{t-1} PO(\hat{k},\mu))(F(t) - F(t-1))}{\sum_t (1 - \sum_{\hat{k}=0}^{t-1} PO(\hat{k},\mu))(F(t) - F(t-1))} \tag{7.4}$$

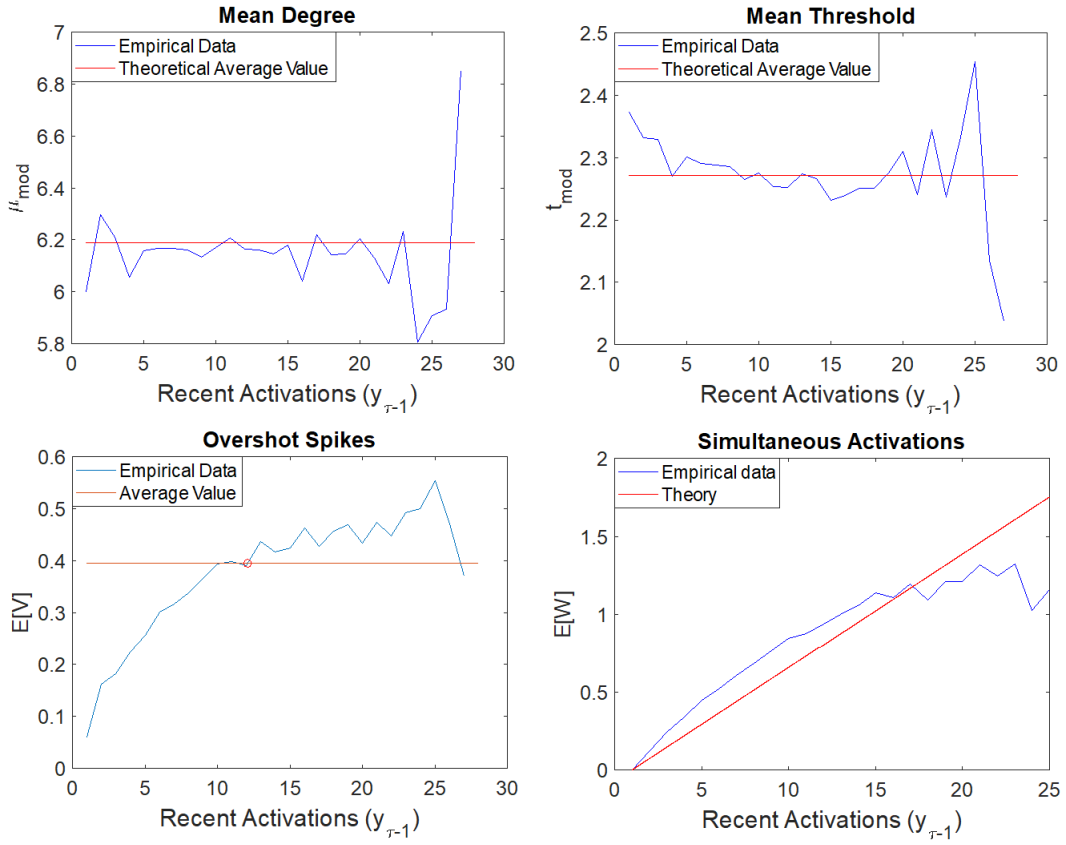

Figure 7.2: Plots comparing the empirical and theoretical values of the average degree $\mu_{\text{mod}}(y_{\tau-1})$ of recently activated agents, the mean threshold $t_{\text{mod}}(y_{\tau-1})$ of those agents, the mean number of overshot spikes $V(y_{\tau-1})$ received by each of the recently activated agents, and $W(y_{\tau-1})$, the mean number of neighbors of each recently activated agent which also activated in the most recent time step.

Again, this parameter is independent of y_τ and independent of τ. We check the validity of this approximation by comparing the empirical value of t_{mod} to the approximation obtained from the equation above. The comparison is plotted in Figure 7.2. While there is some systematic discrepancy between the approximation and the empirical average, this discrepancy is slight.

To develop a theoretical approximation for $V(y_{\tau-1})$, the average number of overshot spikes received by each agent that activated at time $\tau-1$, we need some assumption about the likelihood of an agent activating by exactly reaching its threshold, or exceeding its threshold by any given number of surplus spikes. That is, we want to know the likelihood that an agent of degree k and threshold $t \leq k$ will have any number $t+v$ active neighbors the moment

before it activates, for $0 \leq v \leq k-t$. This probability is a function of k, t, and v. Obviously, this probability does not exist for $t > k$, because then the agent cannot activate, so it is meaningless to discuss the probability of its having any given number of active neighbors when it activates (which will never happen). If an agent has $t > k$ then that agent can be considered immune to the cascade. While we keep this probability small enough not to have a significant effect on the cascade size, there has been study on the effect of too many immune agents in a network [25, 37]. We make the naive assumption that V is a constant, and not affected by $y_{\tau-1}$. After all, those $y_{\tau-1}$ agents cannot directly send overshot spikes to each other, as a time step must elapse between the sender and the recipient activating. Under our assumption, we only need one estimate for V, and can use that same estimate for all values of $y_{\tau-1}$. The mean field simulation yields its own average per-capita number of overshot spikes, which we call V_{MFT}, and is calculated

$$V_{\text{MFT}} = \int_0^w \sum_k \sum_t \sum_{m=0}^{\min(k,t-1)} \sum_{n=t-m}^{k-m} \frac{T_{\tau-2}(k,m,t,x) \times p(n|k,m,x,\tau-1,) \times (m+n-t)dx}{E[Y_\tau]}. \quad (7.5)$$

where $p(n|k,m,x,\tau-1)$ is defined as the likelihood of an agent at x with degree k and m already active neighbors receiving n new active neighbors. Because we are using the mean field approximation, we can calculate $p(n|k,m,x,\tau-1)$ using equations (6.6) and (6.7). We make the naive assumption that $V(y_{\tau-1}) = V_{\text{MFT}}$ for any $y_{\tau-1}$. While our values may vary slightly with varying choices of τ, we choose τ large enough that the wave shape would stabilize. Because the cascade would approach a propagating wave by this point, our value of V_{MFT} only has minimal dependence on the time τ that we choose to evaluate it. Unfortunately, as Figure 7.2 shows, using (7.5) to estimate $V(y_{\tau-1})$ is not accurate. Most notably, according to (7.5) $V(y_{\tau-1})$ has no $y_{\tau-1}$ dependence, but as Figure 7.2 shows, $V(y_{\tau-1})$ depends heavily on $y_{\tau-1}$. This issue is addressed in Section 7.2.

To approximate $W(y_{\tau-1})$, we start with the value W_{MFT} predicted by the mean field theory. To calculate W_{MFT}, we calculate the number of pairs of agents within range of each other that both recently activated, multiply by the probability of two agents within range being adjacent, and divide by $E[Y_\tau]$ to get the per-capita value W. This gets us

$$W_{\text{MFT}} = \frac{\mu \times \int_{R(x_1)} \int_0^w [\rho(x_1, \tau) - \rho(x_1, \tau - 1)] \times [\rho(x_2, \tau) - \rho(x_2, \tau - 1)] dx_1 dx_2}{(2r - 1) \times E[Y_\tau]} \tag{7.6}$$

where $R(x_1)$ is the range of x_1. Next, we ask "What is the probability P_{Conn} that two recently activated agents will be adjacent?" If each recently activated agent is adjacent to an average of W_{MFT} of the $E[Y_\tau] - 1$ other recently activated agents then we get

$$P_{\text{Conn}} = \frac{W_{\text{MFT}}}{E[Y_\tau] - 1}. \tag{7.7}$$

We assume that the likelihood P_{Conn} of a pair of recently activated agents being adjacent is independent of $y_{\tau-1}$ and equal to the value obtained from the mean field approximation. This gives us

$$W(y_{\tau-1}) = (y_{\tau-1} - 1) P_{\text{Conn}} \tag{7.8}$$

for $y_\tau > 0$ and $W(0) = 0$ as a special case.

Figure 7.2 displays the accuracy of these assumptions. What's most interesting are the effects of using theoretical approximations of the number of overshot spikes and the probability of two recently activated agents being connected. The average number of overshot spikes to each agent that activated at time $\tau - 1$ is, in general, dependent on $y_{\tau-1}$, the number of agents that activated at that time. The empirical relationship, is gathered from data taken over 30,000 simulations. $V(y_{\tau-1})$ increases with $y_{\tau-1}$, as opposed to being independent of $y_{\tau-1}$. To accurately predict the number of overshot spikes, we would need a better approximation.

There is a similar error in our formula for the number of simultaneous activations for each recently activated agent, $W(y_{\tau-1})$. The empirical curve shows that $W(y_{\tau-1})$ is a highly nonlinear function of $y_{\tau-1}$, which implies that P_{Conn} depends on $y_{\tau-1}$. We will address this problem in section 7.3.

7.2 Improved Approximation for the Number of Overshot Spikes

Given the number of agents that activated at some time τ, we want to estimate the number of overshot spikes received by all of those agents combined. A relevant question is

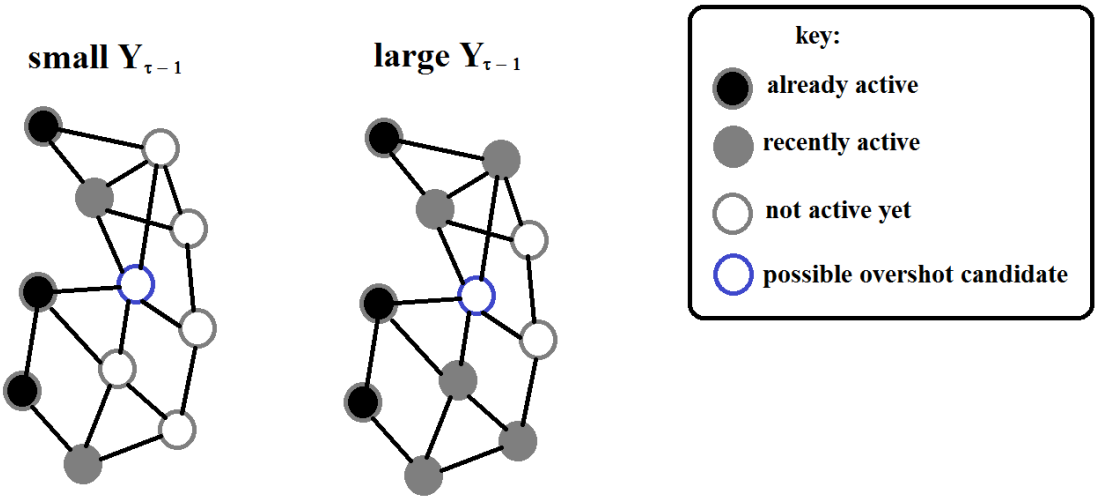

Figure 7.3: A visualization of the relevance of $y_{\tau-1}$ on the number of overshot spikes.

"How many agents activated at time $\tau - 1$?" If that number is particularly high, we would expect more overshot spikes than if it were particularly low, as illustrated in Figure 7.3. In the case where $Y_{\tau-1}$ is particularly high, illustrated in the right panel of the figure, there is a much greater likelihood that an agent would receive multiple spikes between $\tau - 1$ and τ, leading to a greater likelihood of that agent overshooting its threshold. For this reason, we would want to the conditional distribution of $Y_{\tau-1}$ for any given value of Y_τ. This distribution can be approximated from the one-step transition probabilities tabulated in the matrix $\mathbf{P}$ and the quasistationary distribution tabulated in the vector $\mathbf{q}$.

The mean field approximation gives us a single value V_{MFT} (found in (7.9)) of the expected number of overshot spikes received by each of the y_{MFT} agents that activated at time $\tau - 1$ on the mean field network, where y_{MFT} is calculated in (6.12). The point $(y_{\text{MFT}}, V_{\text{MFT}})$ is also plotted in Figure 7.2, and this point falls very close to the empirical curve. Figure 7.4 shows similar comparisons for other response functions, showing that this accuracy is not coincidental. The networks used are similar to our toy network, but the response functions vary. In all cases, $F(2) = F(1)$ and $F(3) = 1$. The values of $F(1)$ are 0.20, 0.35, 0.45, and 0.65. In all four cases, the mean field value falls very close to the empirical curve, giving us a good approximation for one point on the curve. However, as Figure 7.2 shows, $V(y_{\tau-1})$ depends heavily on $y_{\tau-1}$ on the non-cloned network. To understand how to use this statistic, we first ask the question "Which of $y_{\tau-1}$ and y_τ is a better indication of

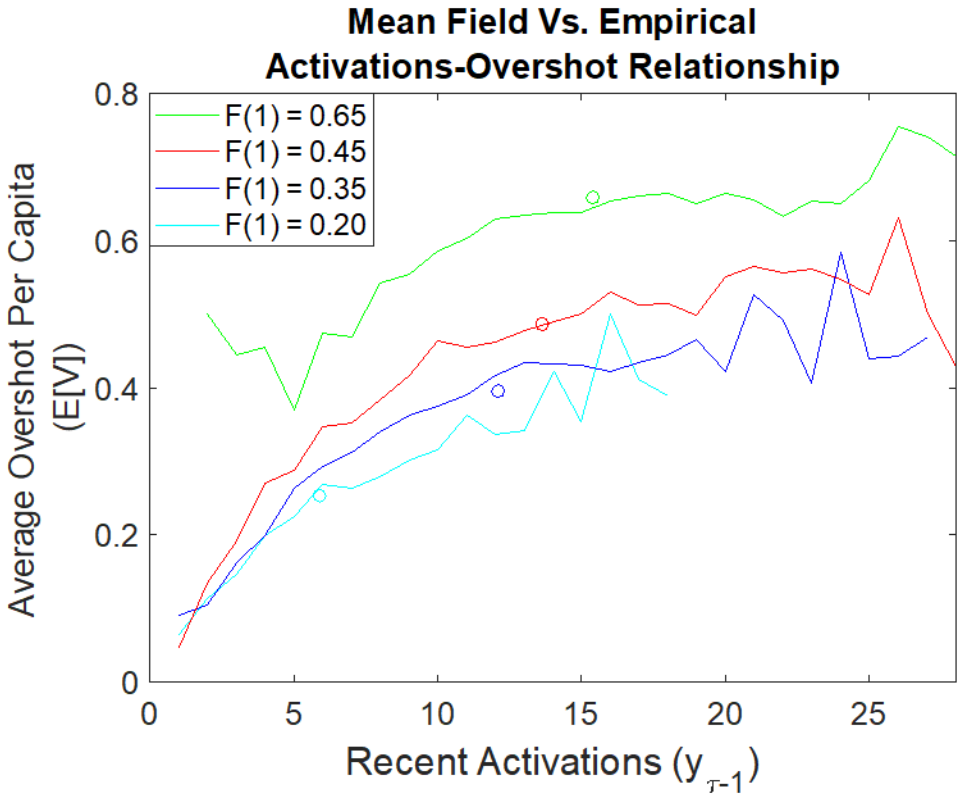

Figure 7.4: Comparisons of the relationship between the number of recent activations $y_{\tau-1}$ and the mean number of overshot spikes received by each of those agents V predicted by the mean field theory (circles) to the empirical curves for networks similar to our toy model, but with varying response functions (line graphs). The empirical data are taken from 10,000 simulations per response threshold distribution.

the total number of overshot spikes sent to the agents that activated at τ?" We propose that they are both equally good indicators. This assumption is vetted in Figure 7.5, where the empirical average number of overshot spikes sent to agents that activated at time 20 is plotted with respect to y_{20} and with respect to y_{19}. The average values closely match each other (barring the oddity at $y_{19} = 24$, which appears to result from a small sample size of cascades with $y_{19} = 24$). For each value of $y_{\tau-1}$ in the mean field calculations we are interested in $\tilde{V}(y_{\tau-1})$, the expected number of overshot spikes induced at the *next* time step by calculating the number of times the $y_{\tau-1}$ agents send overshot spikes to other agents. We assume that $V(y_{\tau-1}) = \tilde{V}(y_{\tau-1})$. That is, we assume that for a given $y_{\tau-1}$, the expected total number of overshot spikes across the network sent by agents that activated at time $\tau - 1$ and received by agents that activated at time τ is the same as the expected number of overshot

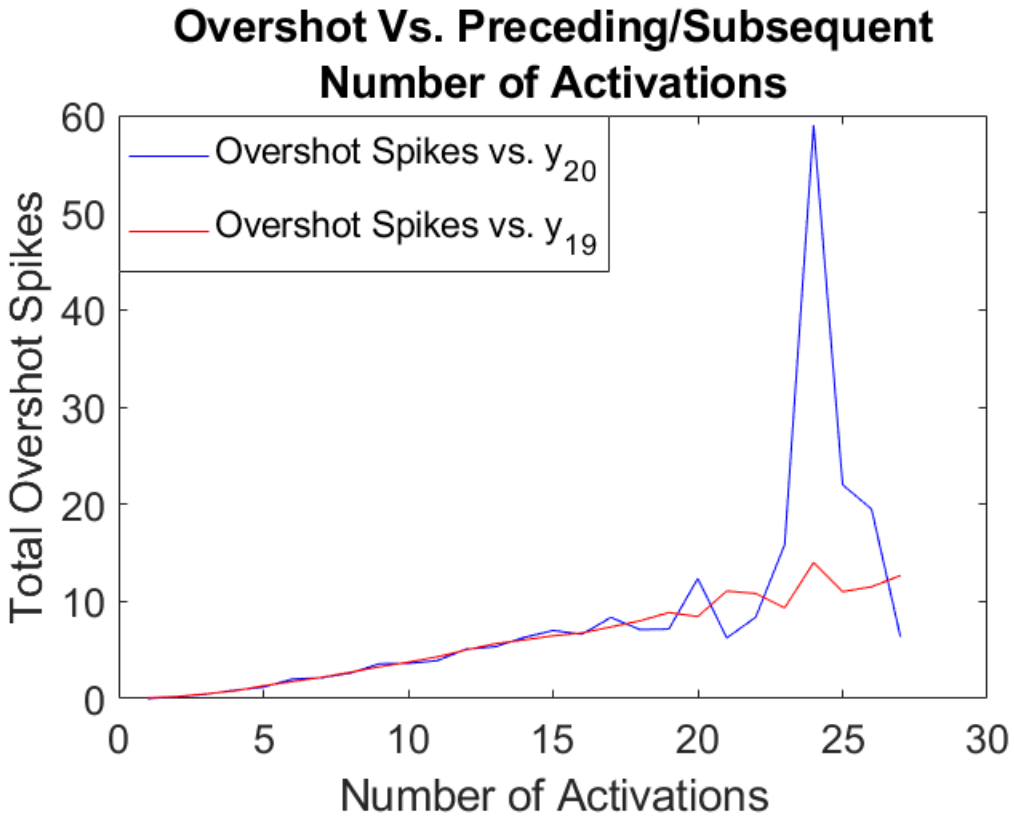

Figure 7.5: A comparison of the empirical average number number of overshot spikes received by the agents that activated at time 20 with respect to the number of preceding activations y_{19} (blue) and the number of induced activations y_{20} (red), taken over 10000 simulations.

spikes sent by agents that activated at time $\tau - 2$ and received by agents that activated at time $\tau - 1$. This assumption is validated by Figure 7.5. We develop the estimate $\tilde{V}_{\mathrm{MFT}}(y_{\tau-1})$ for $\tilde{V}(y_{\tau-1})$ by using the values of $\rho(x, \tau - 2)$ and $\rho(x, \tau - 1)$ on the mean field network with the formula

$$\tilde{V}_{\mathrm{MFT}}(y_{\tau-1}) = \frac{1}{y_{\tau-1}} \int_0^w \sum_k \sum_t \sum_{m=0}^{\min(k,t-1)} \sum_{n=t-m}^{k-m} T_{\tau-1,y_{\tau-1}}(k, m, t, x) \times (n + m - t) \times p(n|k, m, x, \tau, y_{\tau-1}) dx, \quad (7.9)$$

where $p(n|k, m, x, \tau, y_{\tau-1})$ is calculated in (6.26). However, this approach underestimates the number of overshot spikes, as shown in Figure 7.6. As will be discussed later

in Section 7.4, this bias comes from an inaccuracy in the expected distance covered by the wave front between times $\tau - 2$ and $\tau - 1$. Because cloning makes the wave propagate at a faster rate, our predictions based on mean field theory would presume that the agents that activate at time $\tau - 1$ would be further to the right relative to the wave front at time $\tau - 2$ than they actually tend to be on the finite network. This would lead to their being adjacent to more inactive agents and fewer active agents, leading $\tilde{V}_{\text{MFT}}(y_{\tau-1})$ to be an underestimate of $\tilde{V}(y_{\tau-1})$. We will develop a more precise theory for the magnitude of this underestimate in Section 7.4. For now, we use the following simplifying assumption: We know that the mean field analysis gives us a reasonably accurate value for one (noninteger) value of $\frac{y_\tau}{L}$. We multiply $\tilde{V}_{\text{MFT}}(y_{\tau-1})$ by the same factor γ for all $y_{\tau-1}$ so that the interpolated value between $\tilde{V}_{\text{MFT}}(\lfloor E[Y_\tau] \rfloor)$ and $\tilde{V}_{\text{MFT}}(\lceil E[Y_\tau] \rceil)$ matches V_{MFT}. This gets us

$$V_{\text{MFT}} = \gamma \times \\ ((E[Y_\tau] - \lfloor E[Y_\tau] \rfloor) \times \tilde{V}_{\text{MFT}}(\lceil E[Y_\tau] \rceil) + (1 - (E[Y_\tau] - \lfloor E[Y_\tau] \rfloor)) \times \tilde{V}_{\text{MFT}}(\lfloor E[Y_\tau] \rfloor)) \quad (7.10)$$

where V_{MFT} is the average number of overshot spikes sent per agent that activated at time $\tau - 1$ predicted by the mean field theory. This can be rearranged to get

$$\gamma = V_{\text{MFT}} \div \\ ((E[Y_\tau] - \lfloor E[Y_\tau] \rfloor) \times \tilde{V}_{\text{MFT}}(\lceil E[Y_\tau] \rceil) + (1 - (E[Y_\tau] - \lfloor E[Y_\tau] \rfloor)) \times \tilde{V}_{\text{MFT}}(\lfloor E[Y_\tau] \rfloor)). \quad (7.11)$$

With this value of γ, we get

$$\tilde{V}(y_{\tau-1}) = \gamma \times \tilde{V}_{\text{MFT}}(y_{\tau-1}) \quad (7.12)$$

and

$$V(y_{\tau-1}) = \tilde{V}(y_{\tau-1}). \quad (7.13)$$

Figure 7.6 compares the predicted values of $V(y_{\tau-1})$ for each reasonable $y_{\tau-1}$ to its empirical average. This provides a much better estimate than the constant approximation

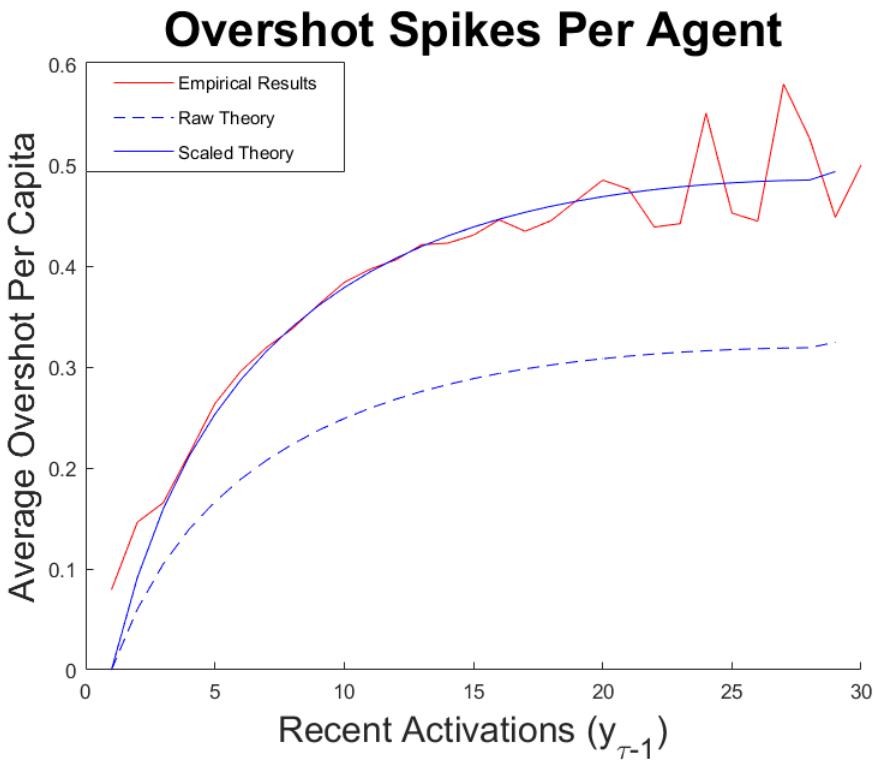

Figure 7.6: A comparison of the empirical average number of overshot spikes received per capita, $V(Y_{\tau-1})$ versus the values generated by our assumptions from section 7.2. While the raw theoretical results are inaccurate, scaling the results to pass through the point (y_{MFT}, V_{MFT}) fixes this issue.

alluded to in section 7.1.

7.3 Improved Approximation for the Number of Simultaneous Activations

When an agent activates, the mean field theory does not accurately predict the expected number of its neighbors which activated simultaneously with itself. This stems from and inaccuracy in the predicted value of $P_{\text{Conn}}(y_{\tau-1})$, the probability of two recently activated agents being connected, given that $y_{\tau-1}$ total agents activated at that time. The agents that activate at time $\tau-1$ have different geographic coordinates, but are likely to be close together because they are not likely to be far from the wave front. On average, agents that activate earlier are likely to be further to the left than those that activate later. If we don't know an agent's location, but we do know when it activated, there is some probability density of

that agent being at any given location. This probability density depends on $\tau - 1$ and the number of agents which activated at that time. On the mean field network, the probability density of the location $X(\tau - 1)$ of agents that activated at time $\tau - 1$ is

$$p_{X(\tau-1)}(x) = \frac{\rho(x, \tau - 1) - \rho(x, \tau - 2)}{\int_0^w \rho(x, \tau - 1) - \rho(x, \tau - 2)dx}, \tag{7.14}$$

where $\rho(x, \tau)$ is the cascade probability function at point x and time τ. We can find the standard deviation $\sigma_{MFT}(y_{\tau-1,\mathrm{MFT}})$ of the locations of the $(y_{\tau-1,\mathrm{MFT}})$ agents that activated at time $\tau - 1$ on the mean field network by

$$\sigma_{MFT}(y_{\tau-1,\mathrm{MFT}}) = \sqrt{\int_0^w x^2 \times p_{X(\tau-1)}(x)dx - (\int_0^w x \times p_{X(\tau-1)}(x)dx)^2}. \tag{7.15}$$

We can also find the number of activations at time $\tau - 1$ on the mean field network $y_{\tau-1,\mathrm{MFT}}$ with the integral

$$y_{\tau-1,\mathrm{MFT}} = \int_0^w \rho(x, \tau - 1) - \rho(x, \tau - 2)dx. \tag{7.16}$$

We run 30,000 simulated cascades on our toy network and find the average sample standard deviation of activation locations $\sigma(y_{\tau-1})$ for each number of activations $y_{\tau-1}$ in the reasonable range $1 < y_{\tau-1} < 30$. ($y_{\tau-1}$ must be strictly greater than 1 for the sample standard deviation to exist.) We take $\tau = 21$, which means that $y_{\tau-1}$ is the number of agents that activated during the twentieth time step. The results of this simulation are plotted in Figure 7.7. We plot the point $(y_{\tau-1,\mathrm{MFT}}, \sigma_{MFT}(y_{\tau-1,\mathrm{MFT}}))$ in relation to the $y_{\tau-1}$-to-$\sigma(y_{\tau-1})$ plot in Figure 7.7. This approximation misses the empirical distribution badly. This would suggest that the variation in the number of new activations changes the spread of where these activations occur. One hypothesis is that the discrepancy primarily stems from the difference between the mean propagation rate on the original network $E[Y_{\tau-1}]$ and the corresponding value predicted by the mean field theory $y_{\tau-1,\mathrm{MFT}}$. This would mean that the wavefront propagation would happen at a speed $\frac{E[Y_{\tau-1}]}{y_{\tau-1,\mathrm{MFT}}}$ times the speed predicted by the mean field approximation. There are many ways that the wave speed can be slowed. Figure 7.8 illustrates some of these alternatives. What seems the most logical is the case of horizontal compression, illustrated in the left panel. Under horizontal compression, the

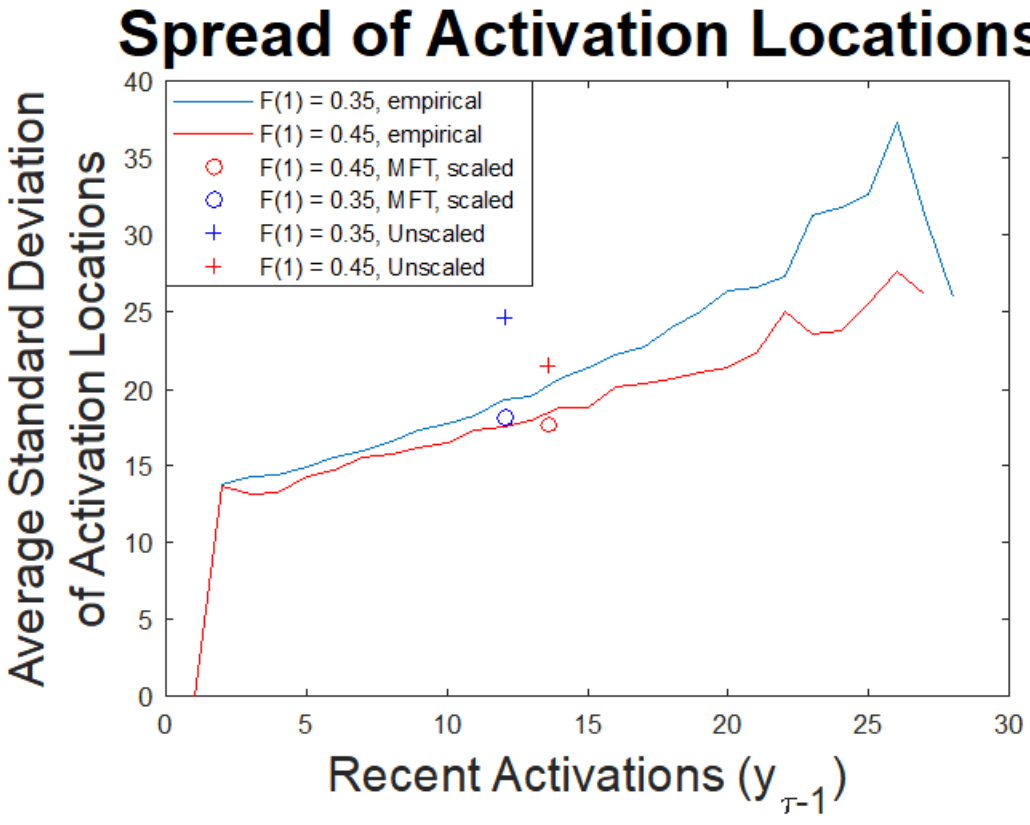

Figure 7.7: The empirical sample standard deviation of the locations of the most recently activated nodes, taken from 30,000 simulated cascades on our toy network, (blue) and a similar network with response threshold distribution $F(1) = 0.45, F(2) = 0.45, F(3) = 1$ (red). These are compared against single-point estimates predicted by the mean field theory on the corresponding networks (+ signs), and the estimates scaled down by the ratio of empirical $E[Y_\tau]$ to the value $E_{\text{MFT}}[Y_\tau]$ predicted by the mean field theory (circles).

distance between any point on the wave at time $\tau-1$ and the corresponding point at time τ is scaled by a factor of $\frac{E[Y_{\tau-1}]}{y_{\tau-1,\text{MFT}}}$. Suppose that the wave passed through (x_1, ρ_1) at time $\tau-1$ and through $(x_1+\epsilon, \rho_1)$ at time τ in the mean field case. On the original network, we would expect the wave to propagate through space more slowly. Suppose that the wave passed through the point (x_2, ρ_2) at time $\tau-1$. We guess that the wave will pass through $(x_2+\frac{E[Y_{\tau-1}]}{y_{\tau-1,\text{MFT}}}\epsilon, \rho_2)$ at time τ. Under this assumption, the standard deviation of locations of activations at time $\tau+1$ would be scaled by that same factor of $\frac{E[Y_{\tau-1}]}{y_{\tau-1,\text{MFT}}}$. This scaling is illustrated in Figure 7.9. As a result, we would scale the expected standard deviation of activation locations by the same factor $\frac{E[Y_{\tau-1}]}{y_{\tau-1,\text{MFT}}}$. Figure 7.10 compares the predicted points $(y_{\tau-1,\text{MFT}}, \sigma_{MFT}(y_{\tau-1,\text{MFT}})\times$

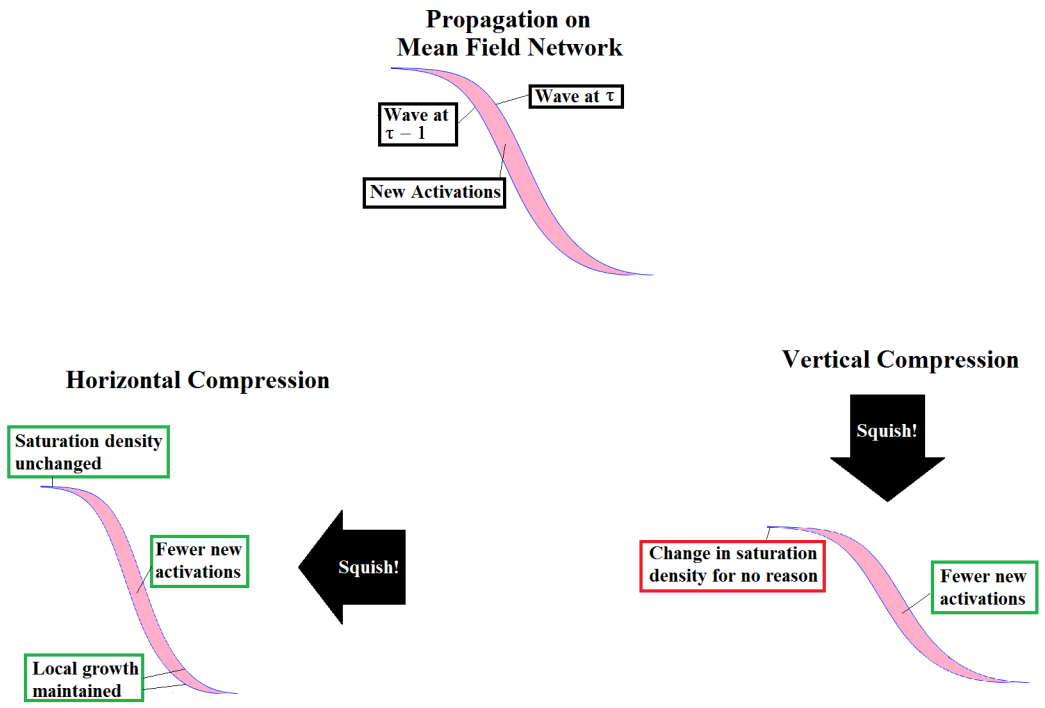

Figure 7.8: Assessment of each of two geometric changes that would decrease $E[Y_{\tau-1}]$. Reasonable and expected implications of the assumptions are outlined in green while unreasonable implications are outlined in red.

$\frac{E[Y_{\tau-1}]}{y_{\tau-1,\mathrm{MFT}}}$) to the empirical distributions for several response threshold distributions. The blue plots correspond to $F(1) = 0.35, F(2) = 0.35, F(3) = 1$ and the red plots correspond to $F(1) = 0.45, F(2) = 0.45, F(3) = 1$. The predicted point is close to the graph for each distribution.

While this provides an accurate approximation of the value of $\sigma(y_{\tau-1})$ for one (noninteger) value of $y_{\tau-1}$, it does not give any indication of the dependence of $\sigma(y_{\tau-1})$ on $y_{\tau-1}$. To determine $\sigma(y_{\tau-1})$ we use the following three quantities:

1) The sample standard deviation of locations of agents that activated at time $\tau - 2$ and sent spikes to agents that activated at time $\tau - 1$. We expect this quantity to depend on $y_{\tau-1}$, so we denote it $\sigma_{\mathrm{dep}}(y_{\tau-1})$.

2) The sample standard deviation of the directed distance between agents that activated at time $\tau - 2$ and sent spikes to agents that activated at time $\tau - 1$ and the agents that received those spikes. We presume this quantity to be independent of $Y_{\tau-1}$ because it is closely related to the radius of influence of the agents. As such, we denote this quantity σ_{ind}.

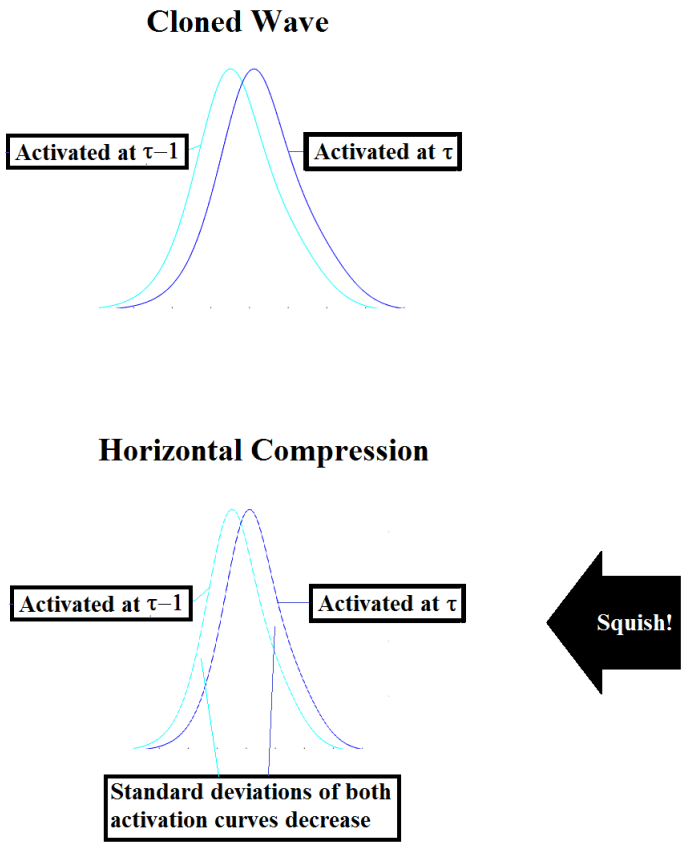

Figure 7.9: Under horizontal compression, the standard deviation of the locations of new activations at any time step decreases.

3) The correlation coefficient ζ between the locations of agents that activated at time $\tau-2$ and sent spikes to agents that activated at time $\tau-1$ and the directed distances between them and the agents that received those spikes. We presume this quantity to be independent of $Y_{\tau-1}$.

Consider an agent that activated at time $\tau-2$ and was adjacent to an agent that activated at time $\tau-1$. Let X_1 be the location of the agent that activated at time $\tau-2$, X_2 be the location of the agent that was inactive at time $\tau-2$ but activated at time $\tau-1$, and $\tilde{X}$ be the directed distance X_2-X_1. We are interested in the bivariate probability density function$f(x_1,\tilde{x})$ of the location of the agent at X_1 and the directed distance between the agent at X_1 and the one at X_2. For the mean field case, we find $f(x_1,\tilde{x})$ given the cascade probability functions $\rho(x,\tau-3)$, and $\rho(x,\tau-2)$. Because the first agent activated at time $\tau-2$ exactly, the probability density of its being at some location x_1 is $\frac{\rho(x_1,\tau-2)-\rho(x_1,\tau-3)}{\int_0^w \rho(x,\tau-2)-\rho(x,\tau-3)dx}$.

Meanwhile, the second agent could be any agent that was not already active at time $\tau - 2$ but active at time $\tau - 1$, so the probability of its being at location x_2 is $\frac{\rho(x_2,\tau-1)-\rho(x_2,\tau-2)}{\int_0^w \rho(x_2,\tau-1)-\rho(x,\tau-2)dx}$. Assuming that the agent that activated at time $\tau - 2$ is as likely to be adjacent to any inactive agent in its range as any other, we get the following formula for $f(x_1, \tilde{x})$:

$$f(x_1,\tilde{x}) = \begin{cases} K(\rho_{\text{new}}(x_1,\tau-2))(\rho_{\text{new}}(x_1+\tilde{x},\tau-1)) & \text{for } x_1+\tilde{x} \in R(x_1) \\ 0 & \text{for } x_1+\tilde{x} \notin R(x_1) \end{cases} \tag{7.17}$$

where K is a normalization constant and $R(x_1)$ is the region within range of influence of x_1. (Because we know that the agents are adjacent, they must be within range of each other, which is why $f(x_1, \tilde{x}) = 0$ for $x_2 \notin R(x_1)$.) For the mean field case, the function $f(x_1, \tilde{x})$ makes it easy to calculate the standard deviation of sender locations $\sigma_{\text{dep,MFT}}$, the standard deviation of the directed distance between the sender and the recipient $\sigma_{\text{ind,MFT}}$, and the correlation coefficient ζ_{MFT} between those two values:

$$\sigma_{\text{dep,MFT}} = \sqrt{\int_{-w}^{w}\int_{0}^{w} {x_1}^2 \times f(x_1,\tilde{x})dx_1 d\tilde{x} - (\int_{-w}^{w}\int_{0}^{w} x_1 \times f(x_1,\tilde{x})dx_1 d\tilde{x})^2}, \tag{7.18}$$

$$\sigma_{\text{ind,MFT}} = \sqrt{\int_{-w}^{w}\int_{0}^{w} \tilde{x}^2 \times f(x_1,\tilde{x})dx_1 d\tilde{x} - (\int_{-w}^{w}\int_{0}^{w} \tilde{x} \times f(x_1,\tilde{x})dx_1 d\tilde{x})^2}, \tag{7.19}$$

and

$$\zeta = \frac{\int_{-w}^{w}\int_{0}^{w} x_1\tilde{x} \times f(x_1,\tilde{x})dx_1 d\tilde{x}}{\sigma_{\text{dep,MFT}} \times \sigma_{\text{ind,MFT}}} - \frac{(\int_{-w}^{w}\int_{0}^{w} \tilde{x} \times f(x_1,\tilde{x})dx_1 d\tilde{x})(\int_{-w}^{w}\int_{0}^{w} x_1 \times f(x_1,\tilde{x})dx_1 d\tilde{x})}{\sigma_{\text{dep,MFT}} \times \sigma_{\text{ind,MFT}}}, \tag{7.20}$$

To get values of σ_{dep}, σ_{ind}, and ζ for the original finite network, we need some assumptions relating σ_{dep}, σ_{ind}, and ζ to their mean field values. Additionally, those three parameters may depend on $y_{\tau-1}$. Because σ_{ind} is largely a product of the constant radius of

influence of each agent, we presume it to be unaffected by cloning or by fluctuations in $Y_{\tau-1}$, giving us

$$\sigma_{\text{ind}}(y_{\tau-1}) = \sigma_{\text{ind,MFT}}. \tag{7.21}$$

We also presume that the correlation coefficient ζ between X_1 and $X_2 - X_1$ is unaffected by cloning or fluctuations in $Y_{\tau-1}$, giving us

$$\zeta(y_{\tau-1}) = \zeta_{MFT}. \tag{7.22}$$

The formula for $\sigma_{\text{dep}}(y_{\tau-1})$ is more complicated. Under our assumption of horizontal compression, $\sigma_{\text{dep}}(y_{\tau-1})$ should scale linearly with the average number of sender agents, $E[Y_{\tau-2}|Y_{\tau-1} = y_{\tau-1}]$. Given the quasistationary distribution $\mathbf{q}$ and the probability transition matrix $\mathbf{P}$ we can use Bayes' theorem to find $P(Y_{\tau-2} = y_{\tau-2}|Y_{\tau-1} = y_{\tau-1})$.

$$P(Y_{\tau-2} = y_{\tau-2}|Y_{\tau-1} = y_{\tau-1}) = \frac{\mathbf{q}_{y_{\tau-2}+1} \times \mathbf{P}_{y_{\tau-2}+1, y_{\tau-1}+1}}{\mathbf{q}_{y_{\tau-1}+1}}. \tag{7.23}$$

Taking the first moment of the above equation gets us $E[Y_{\tau-2}|Y_{\tau-1} = y_{\tau-1}]$. Under our assumption of horizontal compression, σ_{dep} would scale linearly with the expected value of $Y_{\tau-2}$. Note that because the propagation is slower on the finite network than on the mean field network, $E[Y_{\tau-2}]$ is lower than its mean field counterpart $E_{\text{MFT}}[Y_{\tau-2}]$. This gets us the formula

$$\sigma_{\text{dep}}(y_{\tau-1}) = \sigma_{\text{dep,MFT}} \frac{E[Y_{\tau-2}|Y_{\tau-1} = y_{\tau-1}]}{E_{\text{MFT}}[Y_{\tau-2}]}. \tag{7.24}$$

Using the formula for the variance of the sum of two random variables, we have

$$\sigma^2(y_{\tau-1}) = \sigma^2_{\text{ind}}(y_{\tau-1}) + \sigma^2_{\text{dep}}(y_{\tau-1}) + 2\zeta \times \sigma_{\text{ind}}(y_{\tau-1}) \times \sigma_{\text{dep}}(y_{\tau-1}). \tag{7.25}$$

Because we want to approximate P_{Conn}, the likelihood of two agents that activated at time $\tau - 1$ being adjacent, we are interested in how $\sigma(y_{\tau-1})$ affects P_{Conn}. A necessary condition for two agents to be adjacent is that they be within range. We refer to the probability of a given two agents that activated at time $\tau - 1$ being within range of each other, as P_{range}. We can find the probability distribution of the directed distance $\tilde{x}$ between two agents that activated at time $\tau - 1$ on the mean field network with the formula

$$\phi_{\text{MFT}}(\tilde{x}, \tau - 1) = K \int_0^w \rho_{\text{new}}(x_1, \tau - 1) \times \rho_{\text{new}}(x_1 + \tilde{x}, \tau - 2) dx_1. \tag{7.26}$$

where K is a normalization constant. Two agents will be within range if they are separated by no more than the radius of influence r, but not within the same 1-unit interval of the network. We approximate this 1-unit gap in the range of an agent with the gap of radius 0.5 centered around the agent's location. This gives us the following formula for $P_{\text{range,MFT}}$, the probability of two agents that activated at time $\tau - 1$ on the mean field network being within range of each other:

$$P_{\text{range,MFT}} \approx \int_{-r}^{-0.5} \phi_{\text{MFT}}(\tilde{x}, \tau - 1) d\tilde{x} + \int_{0.5}^{r} \phi_{\text{MFT}}(\tilde{x}, \tau - 1) d\tilde{x}. \tag{7.27}$$

The standard deviation of activation locations on the mean field network σ_{MFT} can be found by taking the standard deviation of the activation location X of an agent that activated at time $\tau - 1$, whose probability density is calculated in (7.14). On the original network, if we knew that there were some $y_{\tau-1}$ agents that activated at time $\tau - 1$, we would expect the standard deviation of activation locations to be $\sigma(y_{\tau-1})$ calculated in (7.25) rather than σ_{MFT}. Using the assumption of horizontal compression, the probability of two agents out of the $y_{\tau-1}$ that activated at time $\tau - 1$ being separated by more than some $\tilde{x}$ on the original finite network would be the same as the probability of two such agents being separated by more than $\tilde{x} \times \frac{\sigma_{\text{MFT}}}{\sigma(y_{\tau-1})}$ on the mean field network. This would get us

$$\phi(\tilde{x}, y_{\tau-1}) = \frac{\sigma(y_{\tau-1})}{\sigma_{\text{MFT}}} \times \phi_{\text{MFT}}(\tilde{x} \times \frac{\sigma_{\text{MFT}}}{\sigma(y_{\tau-1})}). \tag{7.28}$$

We can then use the same approximation as the one used in (7.27) to get

$$P_{\text{range}}(y_{\tau-1}) \approx \int_{-r}^{-0.5} \phi(\tilde{x}, y_{\tau-1}) d\tilde{x} + \int_{0.5}^{r} \phi(\tilde{x}, y_{\tau-1}) d\tilde{x}. \tag{7.29}$$

The likelihood of a given pair of agents within range of each other being adjacent is $\frac{\mu}{2r-1}$. If we were to select two agents with likelihood P_{range} of being within range of each other, we would have the following equation for P_{Conn}, the probability that they are adjacent

$$P_{\text{Conn}} = \frac{\mu}{2r - 1} P_{\text{range}}. \tag{7.30}$$

Because both agents activated at time $\tau - 1$, we know that they had enough neighbors

to meet their thresholds. This means that they each have average degree μ_{mod} rather than μ. However, this has no effect on the probability of their being adjacent. Because both agents were inactive at time $\tau - 2$, they could not have influenced each other to activate at time $\tau - 1$. Due to the way that the network is constructed, adjacency to one agent is independent of adjacency to another and adjacency to a specific inactive agent has no effect on the probability of being active.

Combining (7.25) with (7.27) and (7.30), we can find the probability $P_{\text{Conn}}(y_{\tau-1})$ of a pair of agents that activated at time $\tau - 1$ being adjacent, given that $y_{\tau-1}$ total agents activated at that time. Because each of the $y_{\tau-1}$ could be connected to any of the $y_{\tau-1} - 1$ others, we get

$$W(y_{\tau-1}) = (y_{\tau-1} - 1) \times P_{\text{Conn}}(y_{\tau-1}). \tag{7.31}$$

The results of this model are plotted in Figure 7.10. The model closely matches the empirical data, with some slight underestimates of $W(y_{\tau-1})$ for small $y_{\tau-1}$ and slight overestimates of $W(y_{\tau-1})$ for large $y_{\tau-1}$. Having developed reasonably accurate approximations for each type of irrelevant spike, we can use our predicted number of relevant spikes in our prediction for the final cascade size distribution. This adjustment is described in Section 7.4.

7.4 Modifying the Procedure to Account for the Proper Number of Relevant Spikes

Having developed equations (7.2), (7.4), (7.13), and (7.31) as theoretical estimates for the four terms on the right hand side of (7.1), we have a theoretical estimate $U_{\text{theor}}(y_{\tau-1})$ for $U(y_{\tau-1})$, the expected number of relevant spikes sent by each of the $y_{\tau-1}$ agents that activated at time $\tau - 1$.

$$U_{\text{theor}}(y_{\tau-1}) = \mu_{\text{mod}} - t_{\text{mod}} - V(y_{\tau-1}) - W(y_{\tau-1}). \tag{7.32}$$

The method used in section 6.3 can also be used to get a prediction $U_{\text{pred}}(y_{\tau-1})$ of the expected number of relevant spikes sent by each agent that activated at time $\tau - 1$. Each relevant spike sent by agents that activated at time $\tau - 1$ must be received by exactly one agent that was inactive at time $\tau - 1$. By counting the expected number of spikes received

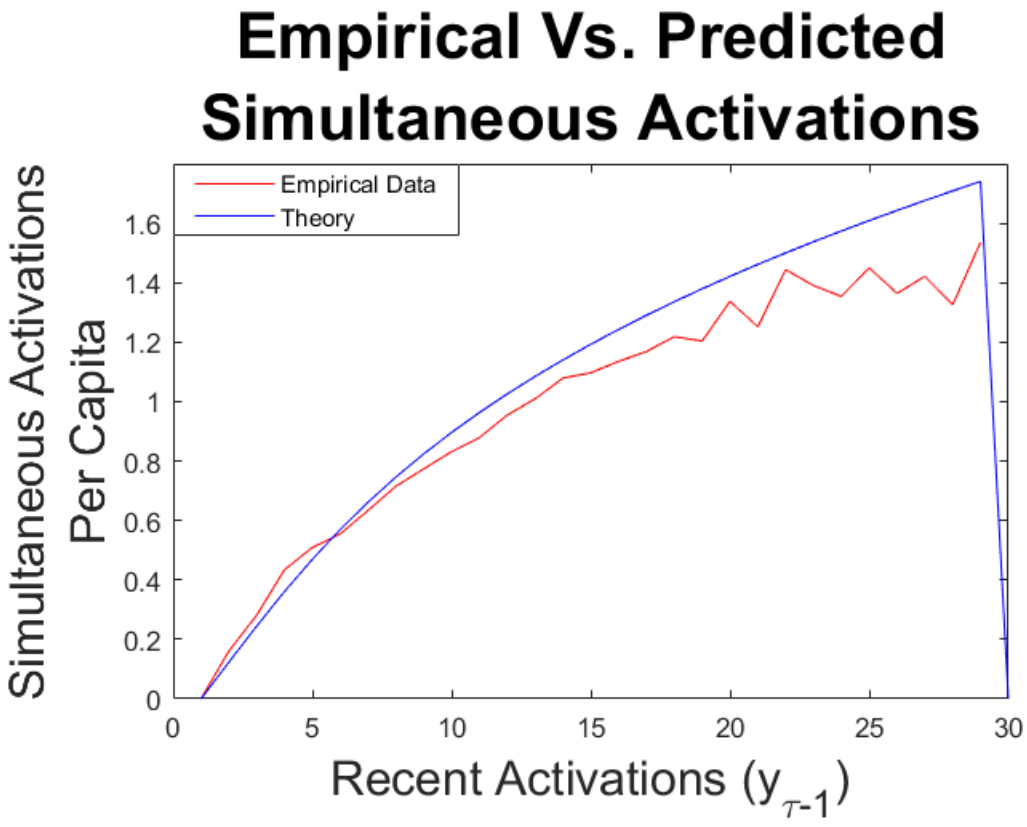

Figure 7.10: Running 30,000 simulated cascades on our toy network, we find empirical estimates for $W(y_{\tau-1})$, the average number of neighbors of each agent that activated at time $\tau - 1$ which activated simultaneously with itself. We compare the empirical data (red) to the values found using the assumptions outlined in this section, (blue).

by all agents that were inactive at time $\tau - 1$ and dividing by the total number of senders $y_{\tau-1}$ we get

$$U_{\text{pred}}(y_{\tau-1}) = \frac{\int_0^w (\sum_k \sum_t \sum_{m=0}^{\min(k,t-1)} \sum_{n=0}^{k-m} F(k,m,t,x,\tau-1,y_{\tau-1},n))dx}{y_{\tau-1}} \tag{7.33}$$

where

$$F(k,m,t,x,\tau-1,y_{\tau-1},n) = T_{\tau-1,y_{\tau-1}}(k,m,t,x) \times p(n|k,m,x,\tau,y_{\tau-1}) \times n), \tag{7.34}$$

w is the width of the network, $p(n|k,m,x,\tau-1,y_{\tau-1})$ is the probability that an agent at location x with degree k and m previously active neighbors will receive n new spikes, given that $y_{\tau-1}$ agents activated at time $\tau-1$, and $T_{\tau-1,y_{\tau-1}}(k,m,t,x)$ is the probability that an agent at point x is inactive at time $\tau-1$ and has degree k, threshold t, and m previously active neighbors. (Because $T_{\tau-1,y_{\tau-1}}(k,m,t,x)$ only accounts for inactive agents, only relevant spikes are counted by (7.33).) The function $p(n|k,m,x,\tau,y_{\tau-1})$ is calculated in (6.26), reiterated here:

$$p(n|k,m,x,y_{\tau-1}) = \frac{\binom{k-m}{n}\binom{N-k-1}{y_{\tau-1}-n}\omega(x,\tau-1)^n}{\sum_{\hat{n}=0}^{min(k-m,y_{\tau-1})}\binom{k-m}{\hat{n}}\binom{N-k-1}{y_{\tau-1}-\hat{n}}\omega(x,\tau-1)^{\hat{n}}}, \tag{7.35}$$

where $\omega(x,\tau-1)$ is the ratio of the activation probability of a not-already-active neighbor of an agent at location x to the activation probability of a non-neighbor of the agent at location x, calculated $\omega(x,\tau-1)=\frac{\omega_1(x,\tau-1)}{\omega_2(x,\tau-1)}$ using (6.22) and (6.25).

Having developed a formula for $U_{\text{pred}}(y_{\tau-1})$, the average number of still inactive agents adjacent to an arbitrary recently activated agent, we develop a similar formula for $W_{\text{pred}}(y_{\tau-1})$ the number of recently activated agents adjacent to a given recently activated agent. If there are N total agents on a network of width w then the linear density of recently activated agents is $\frac{N}{w}(\bar{\rho}_{\text{new},y_{\tau-1}}(x,\tau-1))$, where $\bar{\rho}_{\text{new},y_{\tau-1}}(x,\tau-1)$ is the probability that an agent at location x activated at time $\tau-1$ exactly, given that $y_{\tau-1}$ agents activated overall at time $\tau-1$. If we didn't know whether or not the agent at location x activated at time $\tau-1$ then the expected value of the number of other agents that activated at time $\tau-1$ would be $y_{\tau-1}-\bar{\rho}_{\text{new},y_{\tau-1}}(x,\tau-1)$. If we know that an agent at x activated at time $\tau-1$ then there are only $y_{\tau-1}-1$ other agents that activated at time $\tau-1$. If we know that an agent at location x activated at time $\tau-1$ and that $y_{\tau-1}$ total agents activated at time $\tau-1$ then we approximate the resulting probability that an agent at a different location $\tilde{x}$ also activated at time $\tau-1$ by $\frac{N}{w}\frac{y_{\tau-1}-1}{y_{\tau-1}-\bar{\rho}_{\text{new},y_{\tau-1}}(x,\tau-1)}\bar{\rho}_{\text{new},y_{\tau-1}}(\tilde{x},\tau-1))$. (In reality, this probability would be much more difficult to calculate exactly, relying on Fisher's multivariate noncentral hypergeometric distribution.) Two recently activated agents in range of each other have probability $\frac{\mu}{2r-1}$ of being adjacent. This gets us the formula

$$W_{\text{pred}}(y_{\tau-1}) = \frac{\int_0^w \frac{N}{w}(\bar{\rho}_{\text{new},y_{\tau-1}}(x,\tau-1)\frac{\mu}{2r-1}\int_{R(x)}\frac{N}{w}\frac{y_{\tau-1}-1}{y_{\tau-1}-\bar{\rho}_{\text{new},y_{\tau-1}}(x,\tau-1)}\bar{\rho}_{\text{new},y_{\tau-1}}(\tilde{x},\tau-1))d\tilde{x}dx}{y_\tau - 1}. \quad (7.36)$$

If $U_{\text{pred}}(y_{\tau-1})$ from (7.33) and $U_{\text{theor}}(y_{\tau-1})$ from (7.32) do not match or if $W_{\text{pred}}(y_{\tau-1})$ from (7.36) and $W_{\text{theor}}(y_{\tau-1})$ from (7.31) do not match, we should adjust for this somehow. It should be noted that a discrepancy between the value of $U(y_{\tau-1})$ predicted by our model and the actual value on a real network is significant. If such an error occurs, our model would not correctly predict the number of times an inactive agent had the opportunity to be influenced by a recently activated agent. In contrast, if the model were to predict the right average number of spikes sent from recently activated agents to inactive agents, but have a slight systematic error in determining which inactive agents were the ones to receive those spikes, this would pose a milder problem. It is true that a single inactive agent receiving many spikes is likely to result in only one activation while many agents receiving one spike each would likely result in more than one activation. However, outside of such an extreme case, the effect of a misplaced relevant spike would be much smaller than the effect of a missing or extraneous relevant spike. If we mistakenly assume the agents that activated at time $\tau - 1$ to be too close together or too far apart, this may lead to such an extreme scenario, as it would lead to too many (or too few) inactive agents being within range of many agents that activated at time τ. However, this mistake would also cause $W_{\text{pred}}(y_{\tau-1})$ to differ significantly from $W_{\text{theor}}(y_{\tau-1})$. As long as we adjust our approach to yield $W_{\text{pred}}(y_{\tau-1}) = W_{\text{theor}}(y_{\tau-1})$ and $U_{\text{pred}}(y_{\tau-1}) = U_{\text{theor}}(y_{\tau-1})$ then our prediction for $E[Y_\tau|Y_{\tau-1} = y_{\tau-1}]$ should be reasonably accurate, leading to a reasonably accurate prediction of the CDF of the final cascade size. Sections 7.1 through 7.3 addressed the discrepancy between $U_{\text{pred}}(y_{\tau-1})$ and $U_{\text{theor}}(y_{\tau-1})$ from a bookkeeping perspective. There, we analyzed the average number of neighbors of a given agent that activated at time $\tau - 1$ and estimated the number of those neighbors which had activated before time $\tau - 1$ or at time $\tau - 1$ exactly. Here, we translate that bookkeeping into assumptions on the locations of recently activated agents.

When an agent at location x activates, there are almost exactly $2r - 1$ agents within its range. (Due to the randomness of locations of the agent in each interval, this number ranges from $2r - 2$ to $2r$, with an average value of $2r - 1$.) We want to know how many of

those agents, on average, will be inactive after time $\tau - 1$, activated at time $\tau - 1$ exactly, or active as of time $\tau - 2$. Suppose we have estimates for $\rho_{y_{\tau-1}}(x, \tau - 2)$ and $\rho_{y_{\tau-1}}(x, \tau - 1)$. These can get us estimates of how many recently activated agents and still inactive agents are adjacent to (or within range of) a given recently activated agent.

We compare $W_{\text{pred}}(y_{\tau-1})$ from (7.36) to the value generated by our theory in (7.31) from Section 7.3. We call the later value $W_{\text{theor}}(y_\tau - 1)$. If $W_{\text{pred}}(y_{\tau-1}) > W_{\text{theor}}(y_{\tau-1})$ this would suggest that our method predict the recently activated agents to be too close together. Similarly, if $W_{\text{pred}}(y_{\tau-1}) < W_{\text{theor}}(y_{\tau-1})$, our theory underestimates the spread of the recently activated agents. If $W_{\text{pred}}(y_{\tau-1}) = W_{\text{theor}}(y_{\tau-1})$ but $U_{\text{pred}}(y_{\tau-1})$ and $U_{\text{theor}}(y_{\tau-1})$ are not in agreement, this suggests that while our prediction of the locations of the recently activated agents has them appropriately close to each other, we predict that they are either collectively too far to the left (making them closer to the already activated agents, decreasing $U_{\text{pred}}(y_{\tau-1})$) or too far to the right (which would cause the opposite problem). We need to change $\bar{\rho}_{\text{new},y_{\tau-1}}(x, \tau - 1)$ to a modified function $\hat{\rho}_{\text{new},y_{\tau-1}}(x, \tau - 1)$ fix the issue. One naive approach would be to observe that the errors in the spread and mean location of the agents that activated at time $\tau-1$ correspond to errors in the first two moments of $\bar{\rho}_{\text{new},y_{\tau-1}}(x, \tau-1)$. One way of fixing these issues would be to presume

$$\hat{\rho}_{\text{new},y_{\tau-1}}(x, \tau - 1) = K\bar{\rho}_{\text{new},y_{\tau-1}}(\max(\min(\alpha x + \beta, w), 0), \tau - 1) \tag{7.37}$$

for some α and β and K is a normalization constant. However, this could cause $\hat{\rho}_{\text{new},y_{\tau-1}}(x, \tau - 1) > 1$, so we do not use that assumption.

Instead, we first define the function $\theta(x, \tau - 1)$, which measures the probability that an agent at x will activate at time $\tau - 1$ conditioned on its not already being active, getting us

$$\theta(x, \tau - 1) = \frac{\bar{\rho}_{\text{new}}(x, \tau - 1)}{1 - \rho(x, \tau - 2)}. \tag{7.38}$$

Knowledge of $y_{\tau-1}$ changes the probability that an inactive agent at a given location will activate at time $\tau-1$, which necessitates definition of a modified function $\theta_{\text{mod},y_{\tau-1}}(x, \tau-1)$. Obviously, if more agents overall activated at time $\tau - 1$ then chances are greater that a specific agent will activate at time $\tau - 1$. However, we don't need to account for this in $\theta_{\text{mod},y_{\tau-1}}(x, \tau - 1)$ because we can (and will) use a method similar to that used in Section

6.3 to account for this. Another concern is that, as we have just described, $y_{\tau-1}$ could affect the spread of recently activated agents and the mean location x of recently activated agents. We would like this spreading and shifting to be reflected in $\theta_{\text{mod},y_{\tau-1}}(x, \tau-1)$. Let $\hat{x}$ be the location x where the function $\rho_{\text{new}}(x, \tau-1)$ is maximized. We propose compressing the values of θ about the position $\hat{x}$ by some compression factor $\alpha(y_{\tau-1})$ and then shifting them by some value $\beta(y_{\tau-1})$, to get us

$$\theta_{\text{mod},y_{\tau-1}}(x, \tau-1) = \theta(\min(w, \max(0, \hat{x} + (x - \hat{x} + \beta(y_{\tau-1})) \times (\alpha(y_{\tau-1})))), \tau-1). \quad (7.39)$$

Because $\theta(x, y_{\tau-1}) \approx 0$ anywhere close to the boundary of the network as this region is far from the most active region of the cascade, we can safely impose the minimum of 0 and the maximum of the network width w on the geographic argument of θ. We now use a procedure similar to the one developed in Section 6.3 to find $\bar{\rho}_{y_{\tau-1}}(x, \tau-1)$ and $\bar{\rho}_{y_{\tau-1}}(x, \tau-2)$. The only differences are in the probabilities $P_{\text{Act},y_{\tau-1}}(x, \tau-1)$ from (6.14) and $P_{\text{NotAct},y_{\tau-1}}(x, \tau-1)$ from (6.15). Recall that $P_{\text{Act},y_{\tau-1}}(x, \tau-1)$ is the probability of an agent at x activating at time $\tau-1$ and $y_{\tau-1}$ total agents activating at time $\tau-1$, while $P_{\text{NotAct},y_{\tau-1}}(x, \tau-1)$ is the probability of an agent at x not activating at time $\tau-1$ while $y_{\tau-1}$ total agents activate at time $\tau-1$. These probabilities depend heavily on $\theta_{\text{mod},y_{\tau-1}}(x, \tau-1)$. The equations are now

$$P_{\text{Act},y_{\tau-1}}(x, \tau-1) = (\rho_{\text{sat}} - \rho(x, \tau-2)) \times \theta_{\text{mod},y_{\tau-1}}(x, \tau-1) \times PO(y_{\tau-1}-1, C - (\rho_{\text{sat}} - \rho(x, \tau-2)) \times \theta_{\text{mod},y_{\tau-1}}(x, \tau-1)) \quad (7.40)$$

and

$$P_{\text{NotAct},y_{\tau-1}}(x, \tau-1) = (1 - (\rho_{\text{sat}} - \rho(x, \tau-2)) \times \theta_{\text{mod},y_{\tau-1}}(x, \tau-1)) \times PO(y_{\tau-1}, C - (\rho_{\text{sat}} - \rho(x, \tau-2)) \times \theta_{\text{mod},y_{\tau-1}}(x, \tau-1)) \quad (7.41)$$

where $PO(x, \lambda)$ is the probability that a Poisson random variable with mean λ will have value x and

$$C = \int_0^w (\rho_{\text{sat}} - \rho(x, \tau - 2)) \times \theta_{\text{mod},y_{\tau-1}}(x, \tau - 1)dx. \tag{7.42}$$

We find $\bar{\rho}_{y_{\tau-1}}(x, \tau - 1)$, $\bar{\rho}_{y_{\tau-1}}(x, \tau - 2)$, and $T_{\tau-1,y_{\tau-1}}(k, m, t, x)$ the same way we did in Section 6.3. That is, we find the probability $\rho_{\text{new},y_{\tau-1}}(x, \tau - 1)$ of an agent at location x activating at time $\tau - 1$ exactly given $y_{\tau-1}$ total activations at time $\tau - 1$ with the formula

$$\rho_{\text{new},y_{\tau-1}}(x, \tau - 1) = \frac{P_{\text{Act},y_{\tau-1}}(x, \tau - 1)}{P_{\text{Act},y_{\tau-1}}(x, \tau - 1) + P_{\text{NotAct},y_{\tau-1}}(x, \tau - 1)}, \tag{7.43}$$

where $P_{\text{Act},y_{\tau-1}}(x, \tau - 1)$ and $P_{\text{NotAct},y_{\tau-1}}(x, \tau - 1)$ are calculated in (7.40) and (7.41), which use the function $\theta_{\text{mod},y_{\tau-1}}$. We then normalize the function $\rho_{\text{new},y_{\tau-1}}(x, \tau-1)$ to prevent the paradoxical result

$$\int_0^w \rho_{\text{new},y_{\tau-1}}(x, \tau - 1)dx \neq y_{\tau-1}. \tag{7.44}$$

Normalizing the function $\rho_{\text{new},y_{\tau-1}}(x, \tau - 1)$ gets us

$$\hat{\rho}_{\text{new},y_{\tau-1}}(x, \tau - 1) = \frac{\rho_{\text{new},y_{\tau-1}}(x, \tau - 1) \times y_{\tau-1}}{\int_0^w \rho_{\text{new},y_{\tau-1}}(x, \tau - 1)dx}. \tag{7.45}$$

Again, we would cap $\hat{\rho}_{\text{new},y_{\tau-1}}(x, \tau - 1)$ at 1 if we needed to, but this is not a concern for reasonable values of $y_{\tau-1}$. (Even for the highest value of $y_{\tau-1}$ that we consider, the case $y_{\tau-1} = 29$, $\hat{\rho}_{\text{new},29}(x, \tau - 1)$ never exceeds 0.5.) Using logic similar to that used to develop (6.19) we presume that the likelihood of an agent being active at time $\tau - 2$ given that it does not activate at time $\tau - 1$ exactly gets us

$$\rho_{y_{\tau-1}}(x, \tau - 2) = \frac{\rho(x, \tau - 2)}{1 - \hat{\rho}_{\text{new}}(x, \tau - 1)} \times (1 - \hat{\rho}_{\text{new},y_{\tau-1}}(x, \tau - 1)), \tag{7.46}$$

and the probability that an agent at location x activated at or before time $\tau - 1$ is

$$\rho_{y_{\tau-1}}(x, \tau - 1) = \rho_{y_{\tau-1}}(x, \tau - 2) + \hat{\rho}_{\text{new},y_{\tau-1}}(x, \tau - 1). \tag{7.47}$$

Consider an agent that was inactive at time $\tau - 1$. Its probability of having some degree k, number of already active neighbors m, and threshold t is affected by the fact that it is not already active. However, we assume that this probability is not affected by the total number of agents $y_{\tau-1}$ that activated at time $\tau - 1$. This gives us

$$T_{\tau-1,y_{\tau-1}}(k,m,t,x) = \frac{T_{\tau-1}(k,m,t,x)}{1-\rho(x,\tau-1)} \times (1-\rho_{y_{\tau-1}}(x,\tau-1)). \tag{7.48}$$

Note that the values of $\alpha(y_{\tau-1})$ and $\beta(y_{\tau-1})$ affect the function $\theta_{\text{mod},y_{\tau-1}}(x,\tau-1)$, which affects the functions $\rho_{\text{new},y_{\tau-1}}(x,\tau-1)$ and $\hat{\rho}_{\text{new},y_{\tau-1}}(x,\tau-1)$, which affect the functions $P_{\text{Act},y_{\tau-1}}(x,\tau-1)$ and $P_{\text{NotAct},y_{\tau-1}}(x,\tau-1)$, which affect the functions $\rho_{y_{\tau-1}}(x,\tau-2)$, $\rho_{y_{\tau-1}}(x,\tau-2)$, and $T_{\tau-1,y_{\tau-1}}(k,m,t,x)$. Recall how the functions $U_{\text{pred}}(y_{\tau-1})$ (from (7.33)) and $W_{\text{pred}}(y_{\tau-1})$ (from (7.36)) are calculated. Those equations are reiterated here:

$$W_{\text{pred}}(y_{\tau-1}) = \frac{\int_0^w \frac{N}{w}(\hat{\rho}_{\text{new},y_{\tau-1}}(x,\tau-1)\frac{\mu}{2r-1}\int_{R(x)}\frac{N}{w}\frac{y_{\tau-1}-1}{y_{\tau-1}-\hat{\rho}_{\text{new},y_{\tau-1}}(x,\tau-1)}\hat{\rho}_{\text{new},y_{\tau-1}}(\tilde{x},\tau-1))d\tilde{x}dx}{y_\tau - 1}. \tag{7.49}$$

and

$$U_{\text{pred}}(y_{\tau-1}) = \frac{\int_0^w(\sum_k\sum_t\sum_{m=0}^{\min(k,t-1)}\sum_{n=0}^{k-m}F(k,m,t,x,\tau-1,y_{\tau-1,n}))dx}{y_{\tau-1}} \tag{7.50}$$

where

$$F(k,m,t,x,\tau-1,y_{\tau-1},n) = T_{\tau-1,y_{\tau-1}}(k,m,t,x) \times p(n|k,m,x,\tau,y_{\tau-1}) \times n), \tag{7.51}$$

where $p(n|k,m,x,\tau,y_{\tau-1})$ is calculated with (7.35), reproduced below:

$$p(n|k,m,x,y_{\tau-1}) = \frac{\binom{k-m}{n}\binom{N-k-1}{y_{\tau-1}-n}\omega(x,\tau-1)^n}{\sum_{\hat{n}=0}^{min(k-m,y_{\tau-1})}\binom{k-m}{\hat{n}}\binom{N-k-1}{y_{\tau-1}-\hat{n}}\omega(x,\tau-1)^{\hat{n}}}. \tag{7.52}$$

$\omega(x,\tau-1) = \frac{\omega_1(x,\tau-1)}{\omega_2(x,\tau-1)}$, where $\omega_1(x,\tau-1)$ and $\omega_2(x,\tau-1)$ are calculated using the following minor modifications of (6.22) through (6.25):

$$\omega_1(x,\tau-1) = \frac{\int_{R(x)}\hat{\rho}_{\text{new},y_{\tau-1}}(\hat{x},\tau-1)d\hat{x}}{\int_{R(x)}(1-\rho_{y_{\tau-1}}(\hat{x},\tau-2))d\hat{x}}. \tag{7.53}$$

$$\omega_{2,in}(x,\tau-1)=\frac{\int_{R(x)}\hat{\rho}_{\text{new},y_{\tau-1}}(\hat{x},\tau-1)d\hat{x}}{\int_{R(x)}1d\hat{x}}. \tag{7.54}$$

$$\omega_{2,out}(x,\tau-1)=\frac{\int_0^w\hat{\rho}_{\text{new},y_{\tau-1}}(\hat{x},\tau-1)d\hat{x}-\int_{R(x)}\hat{\rho}_{\text{new},y_{\tau-1}}(\hat{x},\tau-1)d\hat{x}}{\int_0^w 1d\hat{x}-\int_{R(x)}1d\hat{x}}, \tag{7.55}$$

$$\omega_2(x,\tau-1)=\frac{(N-2r)\omega_{2,out}(x,\tau-1)+(2r-k-1)\omega_{2,in}(x,\tau-1)}{N-k-1}. \tag{7.56}$$

The only differences between (7.53) through (7.56) and (6.22) through (6.25) (which calculated the initial functions $\omega_1(x,\tau-1)$ and $\omega_2(x,\tau-1)$) are the use of the new function $\hat{\rho}_{\text{new},y_{\tau-1}}(\hat{x},\tau-1)$ as the probability that an agent at location x will activate at time $\tau-1$ exactly, given that $y_{\tau-1}$ total agents activated at time $\tau-1$, rather than the old function $\bar{\rho}_{\text{new},y_{\tau-1}}(\hat{x},\tau-1)$.

$U_{\text{pred}}(y_{\tau-1})$ and $W_{\text{pred}}(y_{\tau-1})$ depend on $T_{\tau-1,y_{\tau-1}}(k,m,t,x)$ and $\hat{\rho}_{\text{new},y_{\tau-1}}(x,\tau-1)$, which both depend on $\alpha(y_{\tau-1})$ and $\beta(y_{\tau-1})$. We need to find the values of $\alpha(y_{\tau-1})$ and $\beta(y_{\tau-1})$ which yield $W_{\text{pred}}(y_{\tau-1})=W_{\text{theor}}(y_{\tau-1})$ and $U_{\text{pred}}(y_{\tau-1})=U_{\text{theor}}(y_{\tau-1})$, where $W_{\text{theor}}(y_{\tau-1})$ is the value of $W(y_{\tau-1})$ developed in (7.31) and $U_{\text{theor}}(y_{\tau-1})$ is the value of $U(y_{\tau-1})$ developed in (7.32). We use the bivariate secant method developed in [18]. We actually use the bivariate secant method to solve for $\ln(\alpha(y_{\tau-1}))$ and $\beta(y_{\tau-1})$ to keep $\alpha(y_{\tau-1})>0$. We start with the three initial points $\ln(\alpha_0(y_{\tau-1}))=0$, $\beta_0(y_{\tau-1})=0$, $\ln(\alpha_1(y_{\tau-1}))=0.1$, $\beta_1(y_{\tau-1})=0$, and $\ln(\alpha_2(y_{\tau-1}))=0$, $\beta_0(y_{\tau-1})=1$. For each of the ordered pairs $(\ln(\alpha_i(y_{\tau-1})),\beta_i(y_{\tau-1}))$, we find the associated values $U_{\text{pred},i}(y_{\tau-1})$ and $W_{\text{pred},i}(y_{\tau-1})$ from (7.50) and (7.49). We then generate the next guess for $\alpha(y_{\tau-1})$ and $\beta(y_{\tau-1})$ with the matrix equation

$$\begin{bmatrix}\ln(\alpha_{i+1}(y_{\tau-1}))\\ \beta_{i+1}(y_{\tau-1})\end{bmatrix}=\begin{bmatrix}\ln(\alpha_i(y_{\tau-1}))\\ \beta_i(y_{\tau-1})\end{bmatrix}+\mathbf{X}\mathbf{F}^{-1}\begin{bmatrix}U_{\text{theor}}(y_{\tau-1})-U_{\text{pred},i}(y_{\tau-1})\\ W_{\text{theor}}(y_{\tau-1})-W_{\text{pred},i}(y_{\tau-1})\end{bmatrix} \tag{7.57}$$

developed in [18], where

$$\mathbf{X} = \begin{bmatrix} \ln(\alpha_{i-1}(y_{\tau-1})) - \ln(\alpha_{i-2}(y_{\tau-1})) & \ln(\alpha_i(y_{\tau-1})) - \ln(\alpha_{i-1}(y_{\tau-1})) \\ \beta_{i-1}(y_{\tau-1}) - \beta_{i-2}(y_{\tau-1}) & \beta_i(y_{\tau-1}) - \beta_{i-1}(y_{\tau-1}) \end{bmatrix} \tag{7.58}$$

and

$$\mathbf{F} = \begin{bmatrix} U_{\text{pred},i-1}(y_{\tau-1}) - U_{\text{pred},i-2}(y_{\tau-1}) & U_{\text{pred},i}(y_{\tau-1}) - U_{\text{pred},i-1}(y_{\tau-1}) \\ W_{\text{pred},i-1}(y_{\tau-1}) - W_{\text{pred},i-2}(y_{\tau-1}) & W_{\text{pred},i}(y_{\tau-1}) - W_{\text{pred},i-1}(y_{\tau-1}) \end{bmatrix}. \tag{7.59}$$

We iterate the procedure until the Euclidean norm of the error vector $\begin{bmatrix} U_{\text{theor}}(y_{\tau-1}) - U_{\text{pred},i}(y_{\tau-1}) \\ W_{\text{theor}}(y_{\tau-1}) - W_{\text{pred},i}(y_{\tau-1}) \end{bmatrix}$ is less than 10^{-5}.

In the case $y_{\tau-1} = 1$, $W_{\text{theor}}(y_{\tau-1}) = 0$ and $W_{\text{pred}}(y_{\tau-1}) = 0$ because the case of simultaneous activations is impossible. This gives us only one equation when we need two to solve for $\alpha(1)$ and $\beta(1)$. Note that $\alpha(y_{\tau-1})$ has a much stronger effect on $W_{\text{pred}}(y_{\tau-1})$ than on $U_{\text{pred}}(y_{\tau-1})$. Recall that α represents the compression factor of $\theta(x, \tau-1)$, the probability that an agent at location x will activate at time $\tau - 1$ if it did not activate before then. If this compression factor is larger, more agents that activated at time $\tau - 1$ will be adjacent to each other, increasing $W_{\text{pred}}(y_{\tau-1})$. Meanwhile, recall that $\beta(y_{\tau-1})$ represents the right-to-left shift in $\theta(x, \tau - 1)$. An increase in $\beta(y_{\tau-1})$ will cause the recently activated agents to be closer to more already active agents and fewer inactive agents, decreasing $U_{\text{pred}}(y_{\tau-1})$. Meanwhile, as increase in $\alpha(y_{\tau-1})$ will cause more pairs of recently activated agents to be adjacent to each other in most cases, slightly decreasing $U_{\text{pred}}(y_{\tau-1})$, and have a lesser effect on $U_{\text{pred}}(y_{\tau-1})$ otherwise. Put another way, $\alpha(y_{\tau-1})$ has a strong effect on $W_{\text{pred}}(y_{\tau-1})$ and a weak effect on the sum $U_{\text{pred}}(y_{\tau-1}) + W_{\text{pred}}(y_{\tau-1})$ while $\beta(y_{\tau-1})$ has a strong effect on the sum $U_{\text{pred}}(y_{\tau-1}) + W_{\text{pred}}(y_{\tau-1})$ and a weak effect on $W_{\text{pred}}(y_{\tau-1})$. In the case $y_{\tau-1} = 1$, we have $W_{\text{pred}}(1) = 0$, so the sum $U_{\text{pred}}(1) + W_{\text{pred}}(1)$ is just $U_{\text{pred}}(1)$. With the equation $W_{\text{pred}}(1) = 0$ being noninformative, we assume that $U_{\text{pred}}(1)$ and $W_{\text{pred}}(1)$ give us little indication of how big $\alpha(1)$ should be. To fix this, we use linear extrapolation to assume that $\alpha(1) = 2 \times \alpha(2) - \alpha(3)$ and use the secant method to find the value of $\beta(1)$ that results in $U_{\text{theor}}(1) = U_{\text{pred}}(1)$.

Using the modified probability $\hat{\rho}_{\text{new},y_{\tau-1}}(x, \tau - 1)$ of an inactive agent at location x

activating at time $\tau - 1$ exactly given a value of $y_{\tau-1}$ from (7.45) gets us a new value for $\tilde{V}(y_{\tau-1})$ from (7.12), the number of overshot spikes that the $y_{\tau-1}$ agents will induce at time τ. Previously, we used (7.9) to find $\tilde{V}_{\text{MFT}}(y_{\tau-1})$ and multiplied it by the scaling constant γ from (7.11). While we still have

$$\tilde{V}(y_{\tau-1}) = \frac{1}{y_{\tau-1}} \int_0^w \sum_k \sum_t \sum_{m=0}^{\min(k,t-1)} \sum_{n=t-m}^{k-m} T_{\tau-1,y_{\tau-1}}(k,m,t,x) \times (n+m-t) \times p(n|k,m,x,\tau,y_{\tau-1}) dx, \quad (7.60)$$

the values of $T_{\tau-1,y_{\tau-1}}(k,m,t,x)$ and $p(n|k,m,x,\tau,y_{\tau-1})$ now incorporate the compression and shifting factors $\alpha(y_{\tau-1})$ and $\beta(y_{\tau-1})$. This gives us a new set of values $V(y_{\tau-1})$ with the formula

$$V(y_{\tau-1}) = \tilde{V}(y_{\tau-1}). \quad (7.61)$$

Because the values of $V(y_{\tau-1})$ now account for the compression and shifting of the likelihood of an agent at location x activating at time $\tau - 1$, we no longer need to use the scaling factor γ from (7.11), which was intended to fix the systematic underestimate in $V(y_{\tau-1})$. Instead this underestimate is addressed by the compression factor, which predicts a greater likelihood that two agents that activated at time $\tau - 1$ will be close to each other. This increases the probability that two or more agents that activated at time $\tau - 1$ will send spikes to the same inactive agent, which increases $\tilde{V}(y_{\tau-1})$, the expected per capita number of overshot spikes induced in the next time step, which increases $V(y_{\tau-1})$, the expected per capita number of overshot spikes received by the $y_{\tau-1}$ agents that activated at time $\tau - 1$. As will be shown in the left panel of Figure 7.12, the values of $V(y_{\tau-1})$ that we ultimately predict using our method are close to the empirically determined values.

We now calculate the probability transition matrix $\mathbf{P}$ of the number of activations from time $\tau - 1$ to time τ. Recall that the probability $\tilde{\rho}_{\text{new},y_{\tau-1}}(x,\tau)$ of an agent at location x activating at time τ exactly given that $y_{\tau-1}$ agents activated at time $\tau - 1$ is calculated in (6.27), reiterated here:

$$\tilde{\rho}_{\text{new},y_{\tau-1}}(x,\tau) = \sum_k \sum_t \sum_{m=0}^{\min(k,t-1)} \sum_{n=t-m}^{k-m} T_{\tau-1}(k,m,t,x,y_{\tau-1}) \times p(n|k,m,x,\tau,y_{\tau-1}), \quad (7.62)$$

where $p(n|k,m,x,\tau,y_{\tau-1})$ is the probability that an agent at location x that was inactive at time $\tau-1$ and had degree k and m already active neighbors will receive n new spikes at time τ given that $y_{\tau-1}$ total agents activated at time $\tau-1$. $p(n|k,m,x,\tau,y_{\tau-1})$ is calculated in (7.35).

We calculate $E[Y_\tau|Y_{\tau-1}=y_{\tau-1}]$ with (6.28), reiterated here:

$$E[Y_\tau|Y_{\tau-1}=y_{\tau-1}] = \int_0^w \tilde{\rho}_{\text{new},y_{\tau-1}}(x,\tau))dx \quad (7.63)$$

Using the assumption that we can use a Poisson distribution to model the distribution of Y_τ, (validated by Figure 6.7), we have the following equation for the entries in the probability transition matrix $\mathbf{P}$.

$$\mathbf{P}_{y_{\tau-1}+1,y_\tau+1} = p(Y_\tau = y_\tau|Y_{\tau-1}=y_{\tau-1}) = \frac{\lambda^{y_\tau} \times e^{-\lambda}}{y_\tau!} \quad (7.64)$$

where

$$\lambda = \int_0^w \tilde{\rho}_{\text{new},y_{\tau-1}}(x,\tau)dx. \quad (7.65)$$

However, this new probability transition matrix $\mathbf{P}$ necessitates changes to $U_{\text{theor}}(y_{\tau-1})$ and $W_{\text{theor}}(y_{\tau-1})$. Recall that the core of our derivation of $W_{\text{theor}}(y_{\tau-1})$ relied on $\sigma(y_{\tau-1})$, the standard deviation of locations of agents that activated at time $\tau-1$. Recall how $\sigma(y_{\tau-1})$ is calculated in (7.25), reiterated here:

$$\sigma^2(y_{\tau-1}) = \sigma^2_{\text{ind}}(y_{\tau-1}) + \sigma^2_{\text{dep}}(y_{\tau-1}) + 2\zeta \times \sigma_{\text{ind}}(y_{\tau-1}) \times \sigma_{\text{dep}}(y_{\tau-1}), \quad (7.66)$$

where $\sigma_{\text{dep}}(y_{\tau-1})$ is the standard deviation of the locations of agents that activated at time $\tau-2$ and sent spikes to agents that activated at time $\tau-1$, $\sigma_{\text{ind}}(y_{\tau-1})$ is the directed distance between the sender and the recipient, and ζ is the correlation coefficient between $\sigma_{\text{dep}}(y_{\tau-1})$ and $\sigma_{\text{ind}}(y_{\tau-1})$. Because $\sigma_{\text{ind}}(y_{\tau-1})$ is largely a product of the radius of influence of the agents on the network, which has not changed, we assume that $\sigma_{\text{ind}}(y_{\tau-1})$ has the same

value as we predicted in Section 6.3. We also assume ζ to be the same as its value on the mean field network, so it also is unchanged. Recall how $\sigma_{\text{dep}}(y_{\tau-1})$ was calculated in (7.24), reiterated here:

$$\sigma_{\text{dep}}(y_{\tau-1}) = \sigma_{\text{dep,MFT}} \frac{E[Y_{\tau-2}|Y_{\tau-1} = y_{\tau-1}]}{E_{\text{MFT}}[Y_{\tau-2}]}. \tag{7.67}$$

Because we presume the wave to have reached a quasisteady state, we presume that the probability transition matrix $\mathbf{P}$ represents both the probabilities of transition from time $\tau - 1$ to time τ and from time $\tau - 2$ to time $\tau - 1$. With this in mind, Bayes' theorem tells us that

$$P(Y_{\tau-2} = y_{\tau-2}|Y_{\tau-1} = y_{\tau-1}) = \frac{\mathbf{q}_{y_{\tau-2}+1} \times \mathbf{P}_{y_{\tau-2}+1, y_{\tau-1}+1}}{\mathbf{q}_{y_{\tau-1}+1}}, \tag{7.68}$$

where $\mathbf{q}$ represents the distribution of the number of activations at a given time conditioned on the cascade not terminating before then. Because $\mathbf{P}$ changes, $E[Y_{\tau-2}|Y_{\tau-1} = y_{\tau-1}]$ changes in (7.67), which changes $\sigma_{\text{dep}}(y_{\tau-1})$, changing $W_{\text{theor}}(y_{\tau-1})$. More specifically, our formula (7.28) for $\phi(\tilde{x}, y_{\tau-1})$, the probability distribution of the directed distance between two agents that activated at time $\tau - 1$, (reproduced below)

$$\phi(\tilde{x}, y_{\tau-1}) = \frac{\sigma(y_{\tau-1})}{\sigma_{\text{MFT}}} \times \phi_{\text{MFT}}(\tilde{x} \times \frac{\sigma_{\text{MFT}}}{\sigma(y_{\tau-1})}) \tag{7.69}$$

now uses the new value of $\sigma(y_{\tau-1})$. This new value of $\phi(\tilde{x}, y_{\tau-1})$ gets plugged into (7.29) (reproduced below)

$$P_{\text{range}}(y_{\tau-1}) \approx \int_{-r}^{-0.5} \phi(\tilde{x}, y_{\tau-1}) d\tilde{x} + \int_{0.5}^{r} \phi(\tilde{x}, y_{\tau-1}) d\tilde{x}. \tag{7.70}$$

to get a new value of P_{range}, which gets plugged into (7.30) (reiterated below)

$$P_{\text{Conn}} = \frac{\mu}{2r - 1} P_{\text{range}}. \tag{7.71}$$

The resulting value of P_{Conn} gets plugged into (7.31) to get

$$W_{\text{theor}}(y_{\tau-1}) = (y_{\tau-1} - 1) \times P_{\text{Conn}}(y_{\tau-1}). \tag{7.72}$$

Recall from (7.32) that

$$U_{\text{theor}}(y_{\tau-1}) = \mu_{\text{mod}} - t_{\text{mod}} - V(y_{\tau-1}) - W_{\text{theor}}(y_{\tau-1}), \tag{7.73}$$

where μ_{mod} is the mean degree of agents that eventually activate t_{mod} is the mean threshold of agents that activate, $V(y_{\tau-1})$ is the mean number of overshot spikes received by each agent that activated at time $\tau - 1$ and $W_{\text{theor}}(y_{\tau-1})$ is the mean number of neighbors of each of those agents that also activated at time $\tau - 1$. μ_{mod} and t_{mod} are considered constant. As we have just explained, $W_{\text{theor}}(y_{\tau-1})$ changes with α and β, and as (7.61) indicates, $V(y_{\tau-1})$ changes with a change in α and β. As such, we use an iterative procedure to develop a self-consistent theory for $U_{\text{theor}}(y_{\tau-1})$ and $W_{\text{theor}}(y_{\tau-1})$.

To be more specific, a given iteration starts with a set of values of $U_{\text{theor}}(y_{\tau-1})$ from (7.32) and $W_{\text{theor}}(y_{\tau-1})$ from (7.31). We use the bivariate secant method to find the appropriate values of $\alpha(y_{\tau-1})$ and $\beta(y_{\tau-1})$ through iterative use of (7.57). These values of $\alpha(y_{\tau-1})$ and $\beta(y_{\tau-1})$ get us a new set of values for the number of overshot spikes $V(y_{\tau-1})$ from (7.61) and (7.60). While $\alpha(y_{\tau-1})$ and $\beta(y_{\tau-1})$ do not explicitly appear in (7.60), they are embedded in the functions $T_{\tau-1,y_{\tau-1}}(k, m, t, x)$ and $p(n|k, m, x, \tau, y_{\tau-1})$. The values of $\alpha(y_{\tau-1})$ and $\beta(y_{\tau-1})$ also get us a new set of theoretical values $W_{\text{theor}}(y_{\tau-1})$ of the number of simultaneous activation spikes per agent that activated at time $\tau - 1$ through (7.72). While $\alpha(y_{\tau-1})$ and $\beta(y_{\tau-1})$ do not explicitly appear in (7.72), they are embedded in $P_{\text{Conn}}(y_{\tau-1})$. With the new set of values of the number of overshot spikes $V(y_{\tau-1})$ taken from (7.61) and the new set of values of the number of simultaneous activation spikes per agent that activated at time $\tau - 1$ from (7.72) we use (7.73) to get a new set of theoretical values $U_{\text{theor}}(y_{\tau-1})$ of the expected number of relevant spikes sent by each of the $y_{\tau-1}$ agents that activated at time $\tau-1$. This marks the end of that iteration. Each successive iteration starts with the values of $U_{\text{theor}}(y_{\tau-1})$ from (7.73) and $W_{\text{theor}}(y_{\tau-1})$ from (7.72) from the end of the previous iteration. We continue these iterations until the values of $U_{\text{theor}}(y_{\tau-1})$ from (7.73) and $W_{\text{theor}}(y_{\tau-1})$ from (7.72) on successive iterations are within 1% of their corresponding values from one iteration to the next for each reasonable value of $y_{\tau-1}$.

We then use (7.52) from the final iteration to calculate $p(n|k, m, x, \tau, y_{\tau-1})$, the probability of an agent at location x with degree k and m already active neighbors receiving n new spikes from agents that activated at time $\tau - 1$, given that $y_{\tau-1}$ total agents activated at time $\tau - 1$. We use this new formula for $p(n|k, m, x, \tau, y_{\tau-1})$ in (7.62), where we calculate $\tilde{\rho}_{\text{new},y_{\tau-1}}(x, \tau)$, the probability that an agent at location x activated at time τ exactly, given

that there were $y_{\tau-1}$ total activations at time $\tau - 1$ across the whole network. This modified value $\tilde{\rho}_{\text{new},y_{\tau-1}}(x,\tau)$ is used in (7.63) to find the expected number of new activations at time τ given the number $y_{\tau-1}$ of acivations that occurred one time step prior, which is then used in equation (7.64) to calculate the entries in the probability transition matrix $\mathbf{P}$. We find the projected propagation rate and termination probability by using the quasistationary distribution $\mathbf{q}$ of the number of activations per unit time. To find the termination probability, we allow the system to progress from the quasistationary distribution $\mathbf{q}$ through one more time step according to the probability transition matrix $\mathbf{P}$. That is, the termination probability, which we call p_0, is just the first entry of the vector $\mathbf{q}\ \mathbf{P}$. The mean propagation speed $E[Y]$ is the average number of activations per unit time according to the quasistationary distribution. That is, the mean propagation speed can be found with the equation

$$E[Y] = \begin{bmatrix} 0 & 1 & 2 \dots & y_{\max} \end{bmatrix} \mathbf{q}. \tag{7.74}$$

We can then approximate the CDF of the final cascade size $G(z)$ by using (6.34) through (6.36), reiterated here:

$$G_{\text{spon}}(z) = \begin{cases} 0 & \text{for } z < N_0 \\ 1 - (1-p_0)^{\frac{z-N_0}{E[Y]}} & \text{for } z \geq N_0 \end{cases}, \tag{7.75}$$

$$G_{\text{exst}}(z) \approx \begin{cases} 0 & \text{for } z < N_0 \\ \Phi(z, N_0 + (N-N_0)\times\rho_{\text{sat}}, (N-N_0)\times\rho_{\text{sat}}\times(1-\rho_{\text{sat}})) & \text{for } z \geq N_0 \end{cases}, \tag{7.76}$$

and

$$G(z) = 1 - (1 - G_{\text{spon}}(z)) \times (1 - G_{\text{exst}}(z)), \tag{7.77}$$

where $G_{\text{spon}}(z)$ is the probability that the cascade terminates spontaneously, $G_{\text{exst}}(z)$ is the probability that it terminates by saturation, N_0 is the number of initial seeds, ρ_{sat} is the saturation density, and $\Phi(x, \mu, \sigma^2)$ represents the CDF of a Gaussian distribution with mean μ and variance σ^2 as a function of x.

Figure 7.11 plots the predictions used by this approach on three networks, as well

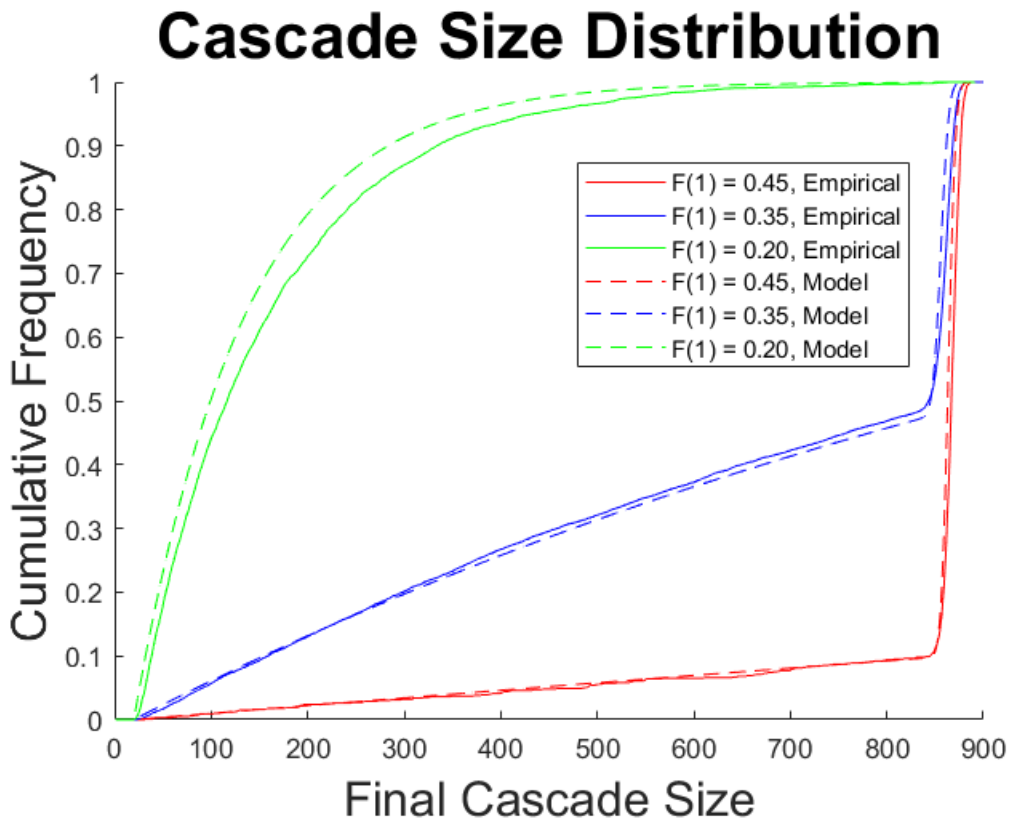

Figure 7.11: A comparison of the final cascade size CDF to the theory. One of the networks used is our toy network. The distributions for this network are plotted in blue. The other networks are similar to our toy network, but have different response threshold distributions. For one network, $F(1) = 0.45, F(2) = 0.45, F(3) = 1$**. The distributions for this response network are plotted in red. For the other network,** $F(1) = 0.20, F(2) = 0.20, F(3) = 1$**. The final cascade size distributions for this network are plotted in green.**

at their empirically determined CDF's taken from 30,000 simulations. One such network is our toy network. The others are similar networks, but with different response threshold distributions. On our toy network, the response threshold distribution is $F(1) = 0.35, F(2) = 0.35, F(3) = 1$. The other networks have response threshold distributions $F(1) = 0.20, F(2) = 0.20, F(3) = 1$ and $F(1) = 0.45, F(2) = 0.45, F(3) = 1$. In each case, the predicted cascade size CDF is close to the empirically determined CDF.

We compare these new values of $V(y_{\tau-1})$ and $W(y_{\tau-1})$ to their corresponding empirical values on our toy network in Figure 7.12. The theory agrees with the empirical data.

As Figure 7.13 shows, for most $y_{\tau-1}$, $\alpha(y_{\tau-1}) > 1$ and $\beta(y_\tau - 1) < 0$ on our toy network.

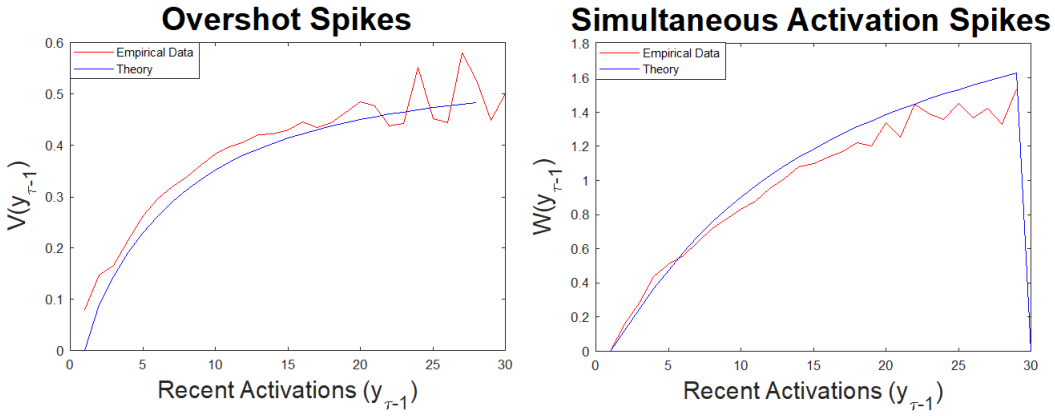

Figure 7.12: A comparison of the expected number of overshot spikes $V(y_{\tau-1})$(left panel) and the expected number of simultaneous activation spikes $W(y_{\tau-1})$ for each reasonable value of the number of activations $y_{\tau-1}$ at time $\tau-1$. The values predicted by our theory are plotted in blue while the empirical data are plotted in red.

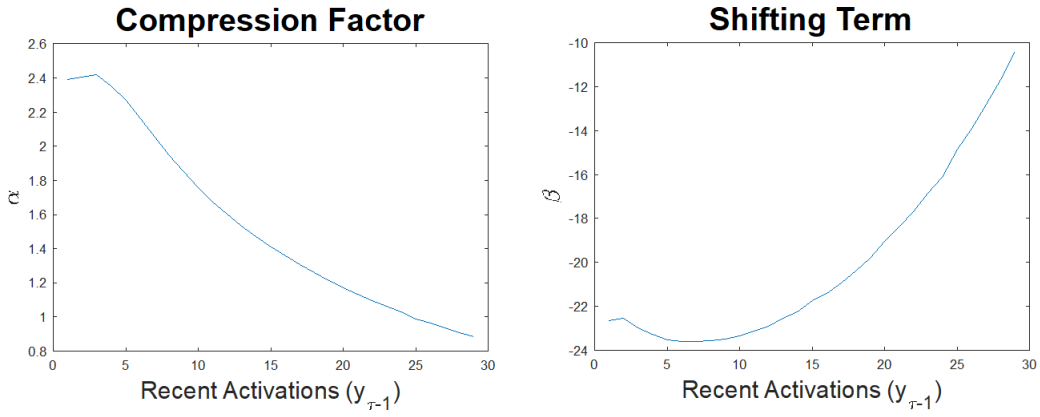

Figure 7.13: Plots showing the compression factor $\alpha(y_{\tau-1})$ (left) and the shifting term $\beta(y_{\tau-1})$ (right) of the probabilities of inactive agents activating at time $\tau-1$ exactly.

This indicates that the agents that activated at time $\tau-1$ are usually closer together and further to the right relative to the time $\tau-2$ wave front than the mean field theory predicts. As Figure 7.14 shows, the values of $U_{\text{pred}}(y_{\tau-1})$ predicted by the methods of Chapter 6 are consistently too small. This means that the agents that activate at time $\tau-1$ should be within range of more agents that were inactive at time $\tau-1$ so that more of the spikes they send are relevant. Because the cascade propagates from left to right, this would push these recently activated agents further to the right. Similarly, as Figure 7.14 shows, the

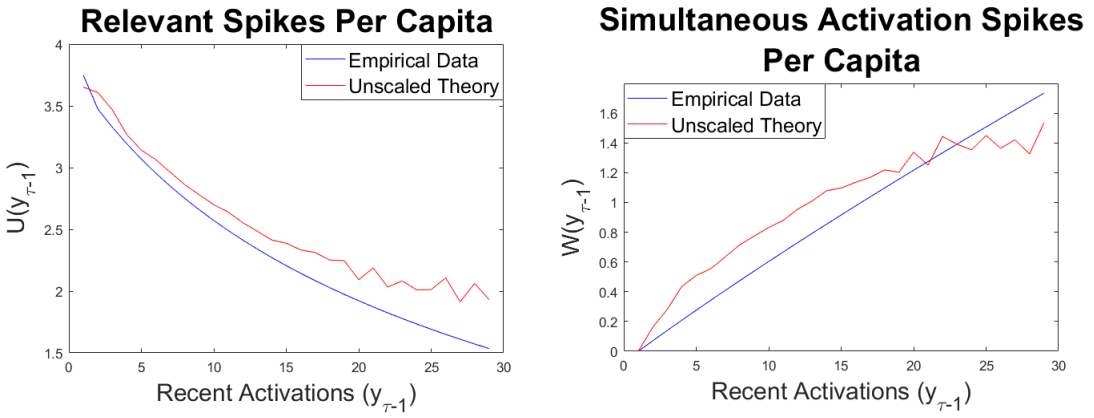

Figure 7.14: Plots showing the number of per-capita relevant spikes sent $U(y_{\tau-1})$ (left) and the per-capita number of simultaneous activations $W(y_{\tau-1})$ (right) on the toy network. The blue curves show the values predicted without compressing or shifting the function $\theta_{y_{\tau-1}}(x, \tau-1)$ (which measures the probability of an agent at location x activating at time $\tau-1$ conditioned on its not activating before then), while the red curves show the empirical average values.

methods of Chapter 6 underestimate the number of spikes sent from agents that activate at time $\tau-1$ to other agents that activated at time $\tau-1$ for most reasonable values of $y_{\tau-1}$. This indicates that the agents that activate at time $\tau-1$ should be closer together than the methods of Chapter 6 indicate. Because the agents that activated at time $\tau-1$ are closer together and further to the right, we can assume that the function $\bar{\rho}_{\text{new},y_{\tau-1}}(x, \tau-1)$, which measures the probability that an agent at point x will activate at time $\tau-1$ given that $y_{\tau-1}$ total agents activated at time $\tau-1$, will also be compressed and shifted to the right to get $\hat{\rho}_{\text{new},y_{\tau-1}}(x, \tau-1)$. This suggests that the function $\theta_{y_{\tau-1}}(x, \tau-1)$, which measure the probability that an agent at location x would activate at time $\tau-1$ if it were inactive before then, should also be compressed and shifted to the right when it is modified to get $\theta_{\text{mod},y_{\tau-1}}(x, \tau-1)$. Given the modification formula (7.39), this implies $\alpha > 1$ and $\beta < 0$.

Note that a shift of β in the function $\theta_{y_{\tau-1}}(x, \tau-1)$ does not lead to a shift of that same β in the peak location of activations at time $\tau-1$. This phenomenon is demonstrated in Figure 7.15. There, for $y_{\tau-1} = 9$, we compare the functions $\theta_{y_{\tau-1}}(x, \tau-1)$ and $\theta_{\text{mod},y_{\tau-1}}(x, \tau-1)$ (plotted in blue) and $\bar{\rho}_{\text{new},y_{\tau-1}}(x, \tau-1)$ and $\hat{\rho}_{\text{new},y_{\tau-1}}(x, \tau-1)$ (plotted in red). The unmodified functions are calculated with $\alpha = 1$ and $\beta = 0$ and plotted in solid curves, and the modified functions are calculated with $\alpha = 1.87$ and $\beta = -23.8$, the values we get on our toy network, as shown by Figure 7.13, and are both plotted in dashed curves. Note that the shift in the

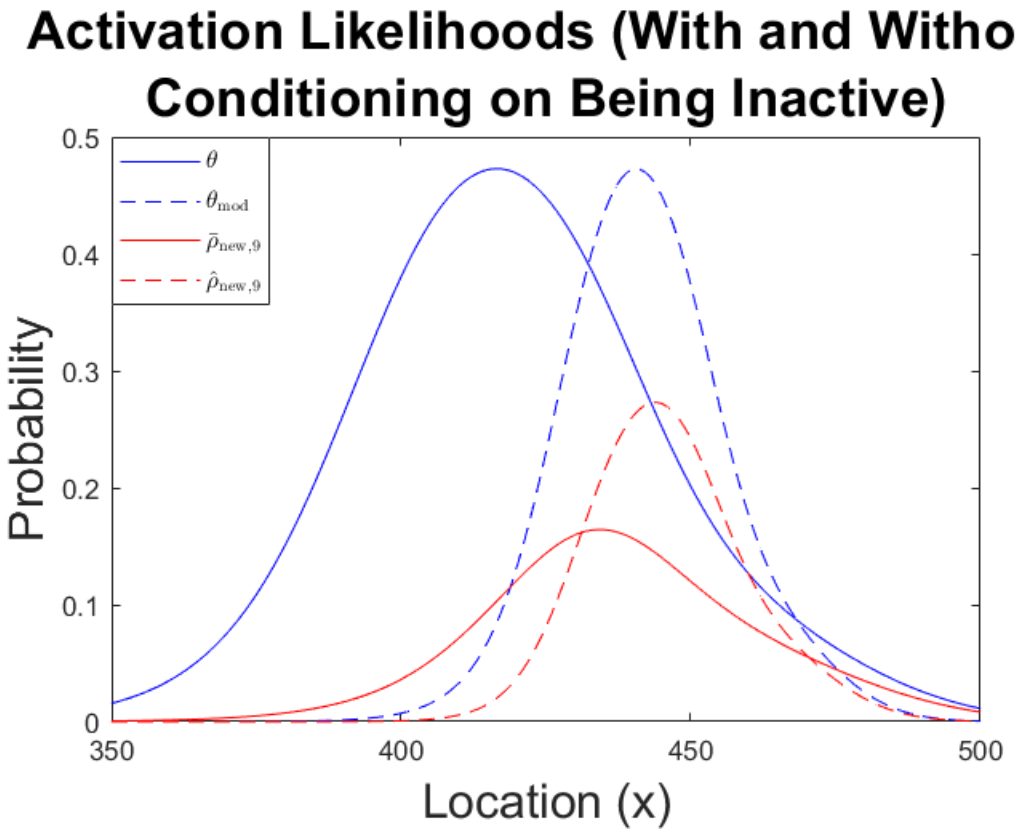

Figure 7.15: A comparison of the probability $\theta_{y_{\tau-1}}(x, \tau-1)$ of an agent at location x activating at time $\tau-1$ conditioned on its not being active at time $\tau-2$ (solid blue curve), to its counterpart $\theta_{\text{mod},y_{\tau-1}}(x, \tau-1)$, which uses the parameters α and β (blue dashed curve), with $y_{\tau-1} = 9$ agents presumed to have activated at time $\tau-1$. The curves are plotted alongside plots of the probabilities $\bar{\rho}_{\text{new},y_{\tau-1}}(x, \tau-1)$ of an agent at location x activating at time $\tau-1$ (solid red curve), and the corresponding modified function $\hat{\rho}_{\text{new},y_{\tau-1}}(x, \tau-1)$, which uses α and β (dashed red curve).

location of maximum $\theta_{y_{\tau-1}}(x, \tau-1)$ is around 24, while the shift in location of maximum $\bar{\rho}_{\text{new},y_{\tau-1}}(x, \tau-1)$ is considerably lower, only around 10. Recall from (7.38) that

$$\hat{\rho}_{\text{new},y_{\tau-1}}(x, \tau-1) = (1 - \rho_{y_{\tau-1}}(x, \tau-2)) \times \theta_{\text{mod},y_{\tau-1}}(x, \tau-1). \tag{7.78}$$

This means that a change in the location of peak time $\tau-1$ activations, which could be found by setting $\frac{d}{dx}\hat{\rho}_{\text{new},y_{\tau-1}}(x, \tau-1) = 0$, would depend not only on the value and x-derivative of $\theta_{\text{mod},y_{\tau-1}}(x, \tau-1)$, but on the value and x-derivative of $(1 - \rho_{y_{\tau-1}}(x, \tau-2))$ as well. Because we don't have closed-form derivatives of the quantities $\theta_{\text{mod},y_{\tau-1}}(x, \tau-1)$

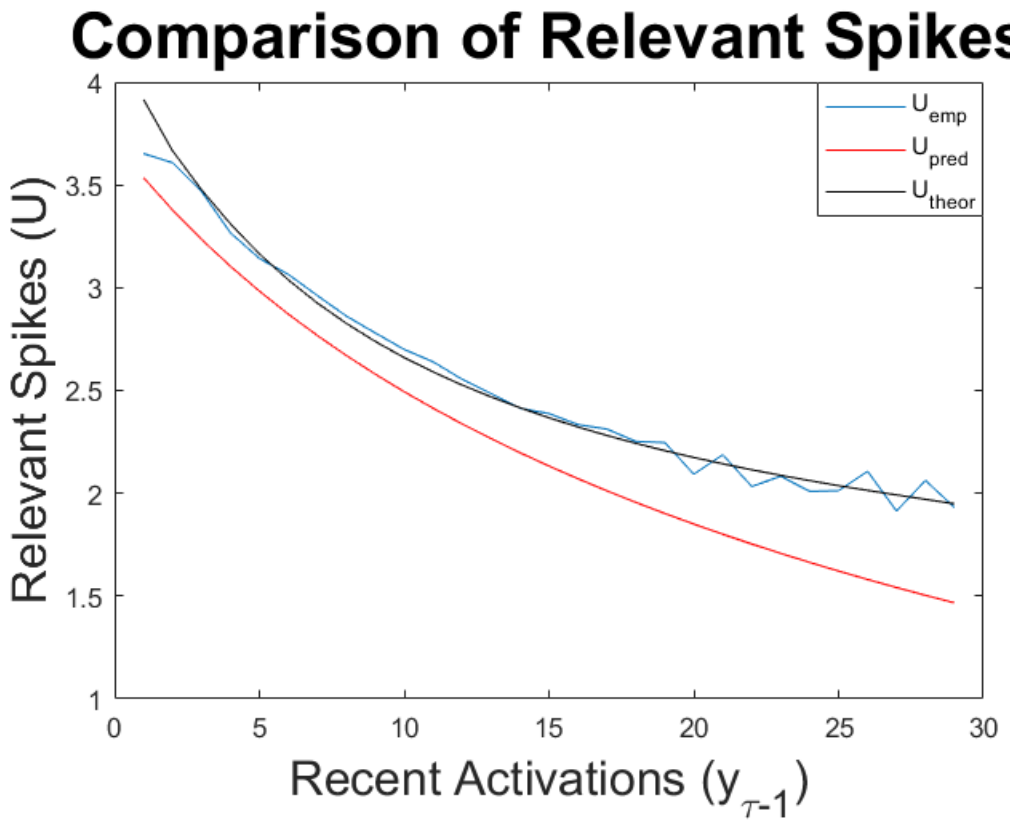

Figure 7.16: A comparison of the empirical number of per-capita relevant spikes, $U_{\text{emp}}(y_{\tau-1})$ (blue) to the values generated by our model, $U_{\text{theor}}(y_{\tau-1})$ (black) and the values $U_{\text{pred}}(y_{\tau-1})$ predicted by the methods on Chapter 6 (red) on our toy network. The empirical data are gathered from 10,000 simulated cascades.

and $(1-\rho_{y_\tau-1}(x,\tau-2))$, we can't get a closed-form equation relating β, the difference in the locations of maximum $\theta_{y_{\tau-1}}(x,\tau-1)$ and $\theta_{\text{mod},y_{\tau-1}}(x,\tau-1)$ to the difference in the locations of maximum $\bar{\rho}_{\text{new},y_{\tau-1}}(x,\tau-1)$ and $\hat{\rho}_{\text{new},y_{\tau-1}}(x,\tau-1)$ found using $\theta_{y_{\tau-1}}(x,\tau-1)$ and $\theta_{\text{mod},y_{\tau-1}}(x,\tau-1)$.

Figure 7.16 plots the empirical values of $U(y_{\tau-1})$, which we call $U_{\text{emp}}(y_{\tau-1})$, taken from 30,000 simulations on our toy network. It also plots the values of $U_{\text{theor}}(y_\tau-1)$, calculated in (7.73). Figure 7.16 also plots the values $U_{\text{pred}}(y_{\tau-1})$ predicted by (7.33). Note that there is better agreement between $U_{\text{emp}}(y_{\tau-1})$ and $U_{\text{theor}}(y_{\tau-1})$ than between $U_{\text{emp}}(y_{\tau-1})$ and $U_{\text{pred}}(y_{\tau-1})$.

CHAPTER 8
CONCLUSIONS

We have developed two successful methods for finding the CDF of the final sizes of cascades on these one-dimensional geographic networks. Each method only requires knowledge of the network statistics. One method, which we developed in Chapter 5, involves the construction of many networks much smaller than the original and finding the cascade sizes on those smaller networks. As Figure 5.7 shows, this approach is very accurate. However, it has the obvious downside that it requires numerous simulations to yield precise results. Because the method uses microcosms of the original network, its complexity does not scale with network size.

The other method only requires a single simulation, and is largely based on analysis of the mean field network. As shown in Figure 7.11 we can accurately predict the final cascade size distribution on one-dimensional geographic cascades. This is more remarkable for network topologies with low probabilities of spontaneous cascade termination. In these cases, a slight error in the mean propagation speed or the mean number of spikes sent from agents that activated at time $\tau - 1$ to agents that were inactive going into time τ could lead to a large relative error in the termination probability. Despite this, the red curves in Figure 7.11 show good agreement between our approximation and the simulated cascade results. We do need to acknowledge the presence of a pair of small canceling errors. As Figure 7.12 shows, we slightly underestimate the number of overshot spikes and slightly overestimate in the number of simultaneous activations. Because these errors are both small, they are of only minor concern. However, because they cancel, the accuracy demonstrated in Figures 7.16 and 7.11 suggests a greater degree of precision than our method actually has.

Obviously, as the network size grows, so does the time it takes for this method to run. However, the increase in complexity is fairly benign. If the local network properties were held constant, but the length were increased, this would hardly affect the algorithm. Once the cascade stabilizes to a wave front propagation, the algorithm only uses the activity in the vicinity of the recently activated agents. If the width of the network were significantly larger than the width of the region with non-negligible activation, the algorithm could be modified to restrict analysis to the part of the network within one radius of influence of the

region with any noticeable increase in $\rho(x, \tau)$, the probability that an agent at location x activated by time τ.

The parameter that presents the largest challenge to running the algorithm quickly is the mean degree. Recall that the ranges of three arguments of the function $T_\tau(k, m, t, x)$ scale with the maximum degree. (The function $T_\tau(k, m, t, x)$ measures the probability than an agent at location x was inactive at time τ and also had k total neighbors, m already active neighbors, and threshold t. The degree k ranges from 0 to the maximum degree, the number of already-active neighbors m ranges from 0 to k, and the threshold t ranges from 0 to $m - 1$. Fortunately, the maximum reasonable degree scales no worse than linearly with the mean degree because the degree distribution is approximately Poisson. Nonetheless, for particularly large networks, it may become necessary to approximate $T_\tau(k, m, t, x)$ by a simplifies $\hat{T}_\tau(k, m, t, x)$. This can be done by binning the entries of $T_\tau(k, m, t, x)$. For example, we may decide to let $\hat{T}(\tilde{k}, \tilde{m}, \tilde{t}, x, \tau)$ represent the likelihood that an agent at point x has degree $2k$ or $2k + 1$, either $2m$ or $2m + 1$ already active neighbors, and response threshold either $2t - 1$ or $2t$. Of course, the rules for binning could vary. As a rule, wider bins will lead to greater efficiency while narrower bins will lead to greater accuracy. For incredibly large networks, with large mean degree and corresponding wide ranges of possible values of m and t, we would expect the values of $T_\tau(k, m, t, x)$ to be similar for similar values of k, m, and t, as long as the restrictions of $m \leq k$ and $m < t$ are followed, so there shouldn't be much error arising from the binning process.

The most noticeable shortcoming of our approach is that it does not generalize well to multiple geographic dimensions. As many real-world networks are embedded on the two-dimensional surface of the Earth, a method would be more pragmatic if it could be applied to two-dimensional geography.

BIBLIOGRAPHY

[1] V.-P. Backlund, J. Saramäki, and R. K. Pan, *Effects of temporal correlations on cascades: Threshold models on temporal networks*, Phys. Rev. E, 89 (2014), p. 062815.

[2] J. P. Bagrow, E. M. Bollt, J. D. Skufca, and D. Ben-Avraham, *Portraits of complex networks*, Europhys. Lett. EPL, 81 (2008), p. 68004.

[3] D. Centola and M. Macy, *Complex contagions and the weakness of long ties*, Amer. J. Soc., 113 (2007), pp. 702–734.

[4] H. Daniels, *The advancing wave in a spatial birth process*, J. Appl. Probab., 14 (1977), pp. 689–701.

[5] J. N. Darroch and E. Seneta, *On quasi-stationary distributions in absorbing discrete-time finite markov chains*, J. Appl. Probab., 2 (1965), pp. 88–100.

[6] M. Del Vicario, A. Bessi, F. Zollo, F. Petroni, A. Scala, G. Caldarelli, H. E. Stanley, and W. Quattrociocchi, *The spreading of misinformation online*, Proc. Nat. Acad. Sci. USA, 113 (2016), pp. 554–559.

[7] W.-B. Du, X.-L. Zhou, Z. Chen, K.-Q. Cai, and X.-B. Cao, *Traffic dynamics on coupled spatial networks*, Chaos Solitons Fractals, 68 (2014), pp. 72–77.

[8] A. Faqeeh, S. Melnik, and J. P. Gleeson, *Network cloning unfolds the effect of clustering on dynamical processes*, Phys. Rev. E, 91 (2015), p. 052807.

[9] J. P. Gleeson, *Cascades on correlated and modular random networks*, Phys. Rev. E, 77 (2008), p. 046117.

[10] ——, *Bond percolation on a class of clustered random networks*, Phys. Rev. E, 80 (2009), p. 036107.

[11] ——, *Binary-state dynamics on complex networks: Pair approximation and beyond*, Phys. Rev. X, 3 (2013), p. 021004.

[12] J. P. Gleeson and D. J. Cahalane, *Seed size strongly affects cascades on random networks*, Phys. Rev. E, 75 (2007), p. 056103.

[13] J. P. Gleeson and S. Melnik, *Analytical results for bond percolation and k-core sizes on clustered networks*, Phys. Rev. E, 80 (2009), p. 046121.

[14] A. Hackett and J. P. Gleeson, *Cascades on clique-based graphs*, Phys. Rev. E, 87 (2013), p. 062801.

[15] A. Hackett, S. Melnik, and J. P. Gleeson, *Cascades on a class of clustered random networks*, Phys. Rev. E, 83 (2011), p. 056107.

[16] Y. Hu, S. Havlin, and H. A. Makse, *Conditions for viral influence spreading through multiplex correlated social networks*, Phys. Rev. X, 4 (2014), pp. 169–179.

[17] W.-M. Huang, L.-J. Zhang, X.-J. Xu, and X. Fu, *Contagion on complex networks with persuasion*, Sci. Rep., 6 (2016), p. 23766.

[18] J. Jankowska, *Multivariate secant method, mathematical models and numerical methods*, Banach Center Publ., 3 (1978), pp. 233–236.

[19] L. Jiang, X. Jin, Y. Xia, B. Ouyang, and D. Wu, *Dynamic behavior of the interaction between epidemics and cascades on heterogeneous networks*, Europhys. Lett. EPL, 108 (2014), p. 58009.

[20] P. D. Karampourniotis, S. Sreenivasan, B. K. Szymanski, and G. Korniss, *The impact of heterogeneous thresholds on social contagion with multiple initiators*, PloS One, 10 (2015), p. e0143020, doi:10.1371/journal.pone.0143020.

[21] F. Karimi and P. Holme, *Threshold model of cascades in empirical temporal networks*, Phys. A, 392 (2013), pp. 3476–3483.

[22] Y. Koç, M. Warnier, P. Van Mieghem, R. E. Kooij, and F. M. Brazier, *The impact of the topology on cascading failures in a power grid model*, Phys. A, 402 (2014), pp. 169–179.

[23] K.-M. Lee, C. D. Brummitt, and K.-I. Goh, *Threshold cascades with response heterogeneity in multiplex networks*, Phys. Rev. E, 90 (2014), p. 062816.

[24] A. Li, X. Zhang, and Y. Pan, *Resistance maximization principle for defending networks against virus attack*, Phys. A, 466 (2017), pp. 211–223.

[25] J. Liu, X. Jin, L. Jiang, Y. Xia, B. Ouyang, F. Dong, Y. Lang, and W. Zhang, *Threshold for the outbreak of cascading failures in degree-degree uncorrelated networks*, Math. Probl. Eng., 2015 (2015), p. 752893.

[26] H. Ma, Y. Zhu, D. Li, S. Li, and W. Wu, *Loyalty improvement beyond the seeds in social networks*, J. Comb. Optim., 29 (2015), pp. 685–700.

[27] S. Melnik, A. Hackett, M. A. Porter, P. J. Mucha, and J. P. Gleeson, *The unreasonable effectiveness of tree-based theory for networks with clustering*, Phys. Rev. E, 83 (2011), p. 036112.

[28] A. Mirshahvalad, A. V. Esquivel, L. Lizana, and M. Rosvall, *Dynamics of interacting information waves in networks*, Phys. Rev. E, 89 (2014), p. 012809.

[29] D. Mollison, *Spatial contact models for ecological and epidemic spread*, J. R. Stat. Soc. Ser. B Stat. Methodol., (1977), pp. 283–326.

[30] K. A. Newhall, M. S. Shkarayev, P. R. Kramer, G. Kovačič, and D. Cai, *Synchrony in stochastically driven neuronal networks with complex topologies*, Phys. Rev. E, 91 (2015), p. 052806.

[31] M. E. Newman, *The structure and function of complex networks*, SIAM Rev., 45 (2003), pp. 167–256.

[32] ——, *Random graphs with clustering*, Phys. Rev. Lett., 103 (2009), p. 058701.

[33] M. E. Newman, S. H. Strogatz, and D. J. Watts, *Random graphs with arbitrary degree distributions and their applications*, Phys. Rev. E, 64 (2001), p. 026118.

[34] J. L. Payne, K. D. Harris, and P. S. Dodds, *Exact solutions for social and biological contagion models on mixed directed and undirected, degree-correlated random networks*, Phys. Rev. E, 84 (2011), p. 016110.

[35] M. A. Porter and J. P. Gleeson, *Dynamical Systems on Networks: A Tutorial*, Springer International, Basel, Switzerland, 2016.

[36] J. Radcliffe and L. Rass, *Discrete time spatial models arising in genetics, evolutionary game theory, and branching processes*, Math. Biosci., 140 (1997), pp. 101–129.

[37] Z. Ruan, G. Iniguez, M. Karsai, and J. Kertész, *Kinetics of social contagion*, Phys. Rev. Lett., 115 (2015), p. 218702.

[38] D. Taylor, F. Klimm, H. A. Harrington, M. Kramár, K. Mischaikow, M. A. Porter, and P. J. Mucha, *Topological data analysis of contagion maps for examining spreading processes on networks*, Nature Commun., 6 (2015), p. 7723.

[39] W. Wang, M. Tang, H.-F. Zhang, and Y.-C. Lai, *Dynamics of social contagions with memory of nonredundant information*, Phys. Rev. E, 92 (2015), p. 012820.

[40] D. J. Watts, *A simple model of global cascades on random networks*, Proc. Nat. Acad. Sci. USA, 99 (2002), pp. 5766–5771.

[41] D. J. Watts and S. H. Strogatz, *Collective dynamics of small-worldnetworks*, Nature, 393 (1998), p. 440.

[42] Z.-K. Zhang, C. Liu, X.-X. Zhan, X. Lu, C.-X. Zhang, and Y.-C. Zhang, *Dynamics of information diffusion and its applications on complex networks*, Phys. Rep., 651 (2016), pp. 1–34.