2013

How Ideas grow: Critical Mass in the Linear Threshold Model

Hossein Alidaee
Macalester College

Honors Project

Macalester College

2013

Title: How Ideas Grow: Critical Mass in the Linear Threshold Model

Author: Hossein Alidaee

MACALESTER COLLEGE

DEPARTMENT OF MATHEMATICS, STATISTICS, AND COMPUTER SCIENCE

How Ideas Grow: Critical Mass in the Linear Threshold Model

Author:
Hossein ALIDAEE

Advisor:
Professor Andrew BEVERIDGE

August 1, 2013

Acknowledgments

Thank you to my readers Dr. Tom Halverson and Dr. Vittorio Addona for generously agreeing to be my second readers. I also thank Dr. Andrew Beveridge and Anna Waggener for a great deal of support throughout the production of this paper. This research was partially supported by the Anderson-Grossheusch Interdisciplinary Summer Research Fund.

Abstract

We study how ideas spread through a social network using the Linear Threshold Model. Each node i on the complete graph K_n is given a threshold θ_i chosen uniformly at random from (0, 1]. This threshold indicates the fraction of the social network that must be active (or believe the idea) prior to node i becoming active. We start with an activated group of early adopters, called the *seed set*. Considering various scenarios, we use the probabilistic method to find lower bounds on size of a seed set which guarantees that all nodes become active with high probability. We characterize seed sets for both homogenous and heterogeneous influence by nodes. In the special case of a single seed node, we draw connections between the Linear Threshold Model and the Catalan numbers.

Contents

Abstract ii

1 Introduction 1

2 Preliminaries 3
- 2.1 Graphs . . . 3
- 2.2 The Linear Threshold Model . . . 4
- 2.3 The Probabilistic Method . . . 8
- 2.4 Asymptotic Analysis . . . 11

3 Inducing Contagion 14
- 3.1 Motivation . . . 14
- 3.2 Research and Methodology . . . 15

4 Contagion in Small Networks 17
- 4.1 Existence of Cohesive Sets . . . 17
- 4.2 Contagion when $|\Phi(0)| = 1$ . . . 25
- 4.3 Contagion when $|\Phi(0)| = k$ . . . 29

5 Contagion in Large Networks 33
- 5.1 A Linear Threshold Model for Asymptotic Analysis . . . 33
- 5.2 Concentrating Bins . . . 34
- 5.3 Inducing Contagion . . . 36

6 Contagion in the Presence of Homophily 39
- 6.1 A Weighted Linear Threshold Model . . . 39
- 6.2 Immutable Thought Leaders . . . 43
- 6.3 Flexible Thought Leaders . . . 46
- 6.4 Alternative Clique Structures . . . 50

7 Future Research 55

Bibliography 59

Chapter 1

Introduction

What determines success of an idea? In recent years, the study of diffusion of innovations has grown from the theory of Everett Rogers, a professor of rural sociology, to a subfield of network science, which studies how the structure of networks affects behavior. Professor Rogers's original manuscript, *Diffusion of Innovations*, sought to explain how, why, and at what rate new ideas, or technologies, are adopted by a population. While his field of study has since expanded its set of tools and research agenda, these three core questions remain at the essence of all its research. In this paper, we study the circumstances under which an entire community will eventually adopt a certain idea or technology.

At first thought, the prospect of an entire community, or *social network*, adopting a technology can be deemed overly ambitious, and even excessive. While this may be true in settings such as markets, where a monopoly is in fact undesirable, there exist certain settings where contagion, adoption by the entire community, is imperative. Perhaps the most illustrative circumstance comes from the field of public health, within the campaign to end malaria.

Malaria, a mosquito-borne infectious disease, is responsible for over 600,000 deaths annually, with approximately 90% taking place in Sub-Saharan Africa (World Malaria Report). Ending malaria is certainly ideal. But how is its eradication relevant to the contagion of technologies in social networks, and why is it dependent on the structure of these networks? Malaria is not a communicable disease, as those traditionally modeled within network science. However, social networks can be linked to the adoption of protective technology, particularly insecticide-treated bed nets (ITNs). This was discovered by Dupas & Cohen (2010), who conducted a randomized field experiment in rural Kenyan villages to determine the effect of vouchers on the use of long-lasting insecticidal nets (LLINs).

LLINs last up to five years in comparison to normal ITNs, which must be replaced every six months. To determine the efficacy of vouchers in adoption of LLINs, Dupas conducted an Ordinary Least Squares regression measuring the effect of witnessing others in the community use this product:

$$Y_{hj1} = \beta High_{hj1} + \gamma ShareHigh_{hj1} + \delta Total_{hj1} + v_j + \epsilon_{hj1}$$

where Y_{hj1} is the random variable indicating whether household h from area j purchased the LLIN during Phase 1 of the trial; $High_{hj1}$ is an indicator variable where the value 1 indicates that the household has received a high subsidy during this phase, v_j is the fixed effect existing for area j, and $Total_{hj1}$ is a variable measuring the number of households within 500 meters to control for reduced likelihood of adoption among those with fewer neighbors. Finally, $ShareHigh_{hj1}$ is our regressor of interest, measuring the effect of the share of neighbors household j has within 500 meters who have received the high subsidy.

Dupas' research found that the share of neighbors who have received the high subsidy, and thus were likely to have redeemed it and have started using the LLIN, had a statistically significant and positive effect in Phase 1 of her trial.[1] However, traditional parametric estimation seems ineffective and unintuitive as a mode of understanding the social effects households experience during technology adoption.

While this OLS regression can provide us with a basic estimation of these effects, it fails to provide a narrative and explanation for household behavior. Further, it assumes that the diffusion of innovation is a linear process, instead of understanding the rich, dynamic changes to the community and their influence on one another as the share of adopters or those provided with the subsidy increases. Finally, such a model fails to answer the core question of many practitioners: what must we do to encourage universal adoption of LLINs in the effort to eradicate malaria?

Network science provides us with an alternative path of exploration. Instead of imposing a linear relationship on a likely nonlinear social system, we model the network by recording which individuals know one another, and attributes both to individuals and their relationships. With this network structure at hand, we can mathematically simulate answers to many of our questions, including the key question behind this paper: how many people must we influence, through non-social efforts, to initially adopt an idea or technology in order for an entire community to eventually undergo adoption as a result of social influence?

[1] Phase 2 of the trial yielded negative effects, believed to be a result of high adoption of these nets and learning. Because LLINs produce the network externality of reduced transmission rates, each individual household's incentive to purchase a net decreases. Further explanation of Phase 2 results can be found in Dupas & Cohen (2010)

Chapter 2

Preliminaries

This chapter provides the mathematical background necessary to understand this paper's main results. A more thorough introduction to the theory of social networks and probabilistic combinatorics can be found in Jackson (2010) and Mitzenmacher & Upfal (2005), respectively.

2.1 Graphs

Before we can model network dynamics, we must address how networks are represented. In a network of n individuals, the set $\mathcal{V} = \{v_1, \ldots, v_n\}$ is the set of **nodes** that are in our network. These nodes can, in the context of social networks, be referred to and thought of as "players" or "individuals." Each of these is indexed by an integer from 1 through n. The canonical form to represent a network of relationships between the set of individuals $\mathcal{V}$ is a graph.

A **graph**, G, consists of a set of nodes $\mathcal{V}(G)$ and a set of edges, $\mathcal{E}(G)$. An edge $e \in \mathcal{E}(G)$ is a 2-element subset of $\mathcal{V}(G)$ indicating that a relationship exists between those two individuals. Each edge is represented visually by a line drawn between two nodes. If the edges of a graph consist of unordered pairs, such that the edge (v_i, v_j) is equivalent to (v_j, v_i), then the graph can be described as **undirected**. Further, if the undirected graph also contains no self-loops and is limited to a single edge per pair of vertices, it considered **simple**. Because our aim is to study the diffusion of information in social networks, we focus on simple graphs, as self-loops and multiple edges are not relevant in this context. Figure 2.1 shows a simple graph representing a social network of 5 people.

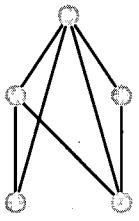


Figure 2.1: A simple graph representing a five person social network.

If an edge exists between each of the n nodes in the network, implying a relationship between each pair of individuals, the graph is considered **complete** and is denoted by K_n. A complete, simple graph representing a five person social network is shown in Figure 2.2.

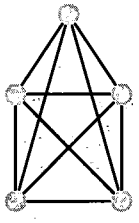


Figure 2.2: The complete, simple graph K_5.

Within a graph $G(\mathcal{V}, \mathcal{E})$, a **subgraph** H is a graph where $\mathcal{V}(H) \subseteq \mathcal{V}(G)$ and $\mathcal{E}(H) \subseteq \mathcal{E}(G)$.

For each individual $v_i \in \mathcal{V}(G)$, we define its neighbors as $\mathcal{N}_i(G) = \{v_j | (v_j, v_i) \in \mathcal{E}(G)\}$. This can be intuitively interpreted as the set of individuals who know and have a relationship with an individual i within the graph G. We indicate the **degree** of each individual —the number of neighbors she has —by $d_i = |\mathcal{N}_i(G)|$.

2.2 The Linear Threshold Model

There are various models for studying the diffusion of ideas or technologies on social networks, known as models of informational cascades. One of the most common informational cascade models is the deterministic **Linear Threshold Model**. In this discrete time model, each individual i is considered to be in one of two states. The first state, **inactive**, indicates they have yet to adopt the behavior. The second state, **active**, indicates they have adopted the behavior. The active state is considered an **absorbing state**, meaning that once an individual is active, they remain active.

To understand how an individual becomes active, we first define their threshold function. For each individual i, this function returns the fraction of her neighbors who must be activated for her to activate as well.

Definition 2.2.1 (Threshold Function). For a graph $G(\mathcal{V}, \mathcal{E})$, the **threshold function** θ is a mapping from $\mathcal{V}$ to $(0, 1]^n$. Formally,

$$\theta : \mathcal{V} \to (0, 1]^n. \tag{2.1}$$

This implies that the number of neighbors that must be active for each individual i is defined by $\kappa_i = \lceil \theta_i \cdot d_i \rceil$. Consequently, we can define a process for activating any individual node.

Definition 2.2.2. Given our set $\Phi(t)$ of active individuals at time t, node v_i is active at time $t+1$ if the number of her active neighbors exceeds her threshold:

$$|\Phi(t) \cap \mathcal{N}_i(G)| \geq \kappa_i \Rightarrow v_i \in \Phi(t+1). \tag{2.2}$$

This process leads to an activation function. For an initial set of active nodes in the network, this returns the set of nodes activated in the next time period.

Definition 2.2.3 (Activation Function). Given a threshold function θ, an **activation function** A_θ is the following mapping:

$$A_\theta(\Phi(t)) = \Phi(t+1) = \Phi(t) \cup \{v_i \in V : |\Phi(t) \cap \mathcal{N}_i(G)| \geq \kappa_i\}. \tag{2.3}$$

It immediately follows that the process has converged when $\Phi(t) = \Phi(t+1)$. To initiate diffusion, there is some subset $\Phi(0) \subseteq \mathcal{V}$ active at time $t = 0$ before the process has begun. This subset is known as the **seed**. Clearly, once $\Phi(t) = \mathcal{V}$, the entire network has adopted the behavior. If a behavior will be adopted by the entire network, it is considered **contagious**. **Contagion** occurs if a behavior is contagious.

Example 2.2.4. Let our social network be represented by the complete graph K_5. Each player's threshold is chosen uniformly at random, leading to the threshold vector $\theta = [0.05, 0.7, 0.6, 0.1, 0.4]$ and thus $\kappa = [1, 3, 3, 1, 2]$. The following figure represents this network, where each player v_i is represented by their value of κ_i.

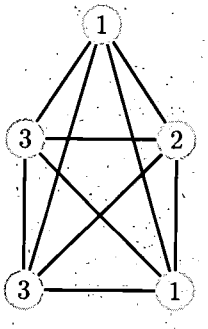


Figure 2.3: A social network represented by K_5.

Suppose that player v_2 is designated as the seed at the beginning of the process. Diffusion proceeds as follows:

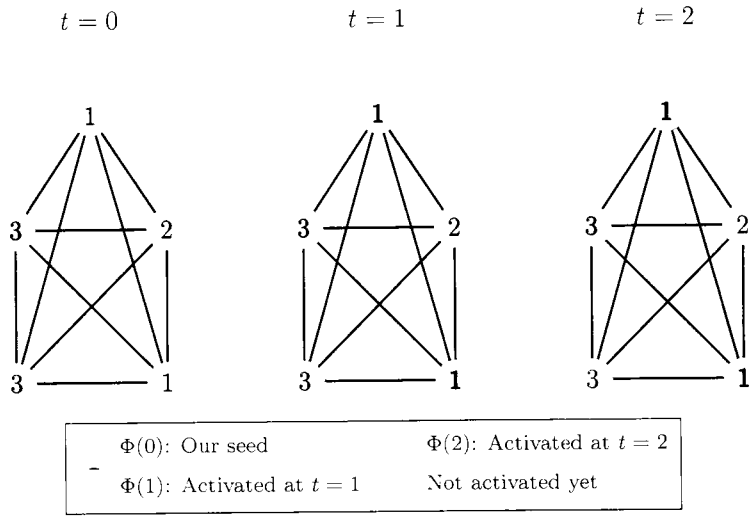


Figure 2.4: The diffusion process for a network on K_5

Diffusion therefore converges and the modeled behavior has been adopted by the entire network by the end of $t = 2$. We can therefore consider the behavior contagious.

Behavior is not always contagious. To understand these other cases, we need several definitions to characterize our social networks. The first definition measures the cohesion of social groups in a network.

Definition 2.2.5 (Cohesive Set)**.** For a graph $G(\mathcal{V}, \mathcal{E})$, a subset of players $\mathcal{M} \subseteq \mathcal{V}$ is a **cohesive set** if

$$\frac{|\mathcal{M} \cap \mathcal{N}_i(G)|}{|\mathcal{N}_i(G)|} > 1 - \theta_i \text{ for all } i \in \mathcal{M}. \tag{2.4}$$

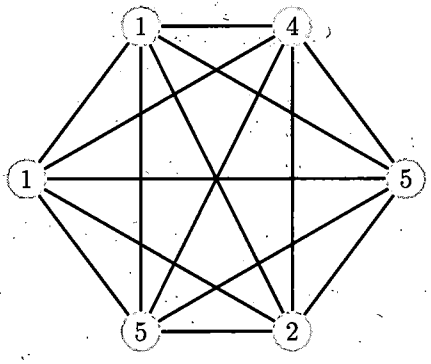
We assume, for completeness, that $\emptyset$, the empty set, is a cohesive set. In all graphs, $\emptyset$ and $\mathcal{V}$ form the trivial cohesive sets.

Definition 2.2.5 states that a group of individuals form a cohesive set if, for each member, her fraction of neighbors not in the set is less than her threshold. Further, it implies that multiple cohesive sets can exist in the network and that these sets must be nested, with each smaller set only including those individuals with higher thresholds. The following examples illustrate cohesive sets in various networks.

Example 2.2.6. In the network provided by Example 2.2.4, there are only two cohesive sets. The first is the entire network, meaning $\mathcal{M} = \{v_1, v_2, v_3, v_4, v_5\}$,

meeting the requirement $\theta_i > 0$ for all $v_i \in \mathcal{M}$. The second cohesive set is the empty set $\emptyset$. We note that these are the two trivial cohesive sets, and exist within any network.

Example 2.2.7. Let our social network be represented by the complete graph K_6. Each player's threshold is chosen uniformly at random, leading to the threshold vector $\theta = [0.05, 0.7, 0.9, 0.25, 0.95, 0.1]$ and thus $\kappa = [1, 4, 5, 2, 5, 1]$. The following figure represents this network, where each player i is represented by their value of κ_i.



Aside from our two trivial cohesive sets (the entire network and the empty set), this network also features other cohesive sets. Each player in the cohesive set $\mathcal{M} = \{v_2, v_3, v_5\}$ has a threshold exceeding the requisite 0.6. Similarly, both players in $\mathcal{M} = \{v_3, v_5\}$ have thresholds exceeding 0.8. Finally, we observe that the nesting property of cohesive sets is held for our network:

$$\{v_1, v_2, v_3, v_4, v_5, v_6\} \supset \{v_2, v_3, v_5\} \supset \{v_3, v_5\} \supset \emptyset.$$

By construction, members of a cohesive set $\mathcal{M}$ cannot satisfy equation (2.2) unless a member of $\mathcal{M}$ has already been activated. This leads to an important characterization of the final set of activated nodes. However, we first need to introduce the concept of a fixed point.

Definition 2.2.8 (Fixed Point)**.** Given a graph $G(\mathcal{V}, \mathcal{E})$ and threshold values θ, a **fixed point** is a nonempty seed set Φ^* for which

$$\Phi^* = \Phi(t) \text{ for all } t \geq 0. \tag{2.5}$$

Lemma 2.2.9 (Acemoglu et al. (2011))**.** *Given a graph $G(\mathcal{V}, \mathcal{E})$ and threshold values θ, a seed Φ^* is a fixed point if and only if $(\Phi^*)^C = \mathcal{V} \setminus \Phi^*$ is a cohesive set.*

The proof of this is based on the aforementioned idea that a member of the cohesive set $\mathcal{M}$ cannot satisfy equation (2.2) unless some other member of $\mathcal{M}$ has already been activated. Therefore, if diffusion does not begin with a member of the seed in a cohesive set, no member of that set will ever become active. The

social network in Example 2.2.7 has several fixed points, each a complement of a cohesive set: $\{v_1, v_4, v_6\}, \{v_1, v_2, v_4, v_6\}, \{v_1, v_2, v_3, v_4, v_5, v_6\}$. Cohesive sets and fixed points provide a vital framework for understanding whether behavior is contagious. A more elaborate discussion can be found in Acemoglu et al. (2011).

2.3 The Probabilistic Method

The probabilistic method is a powerful tool in discrete mathematics. The method proves the existence of mathematical structures with certain desired properties. To do this, it defines a probability space of structures in which the probability that a randomly selected structure has the required properties is positive. As an example, if we can show that there is a positive probability of a student having blonde hair and blue eyes, then there must be at least one student with blonde hair and blue eyes.

Once the thresholds are known, the Linear Threshold Model is deterministic. While some approaches to studying diffusion preconceive all thresholds as a value such as 1/2 (Morris, 2000; Berger, 2001), our research instead assumes that the threshold values are unknown. We select θ uniformly at random to reflect our lack of knowledge about its value. By drawing θ from a probability distribution, the tools of the probabilistic method become necessary.

While the expectations of these distributions are fairly simple to calculate, we often need to focus on the **tail distribution**, the probability that a random variable assumes values that are far from its expectation. In the context of determining whether a behavior is contagious, bounding the tail distribution is important. This bounding procedure is known as the second moment method. The principal tool for this method is Chebyshev's Inequality.

Theorem 2.3.1 (Chebyshev's Inequality). *Let X be a random variable with finite expected value μ and finite, non-zero variance σ^2. Then, for any real number $\lambda > 0$,*

$$\Pr(|X - \mu| \geq \lambda\sigma) \leq \frac{1}{\lambda^2}.$$

This inequality is most relevant when an individual is wondering what the percentage chance is of an observation lying within λ standard deviations of the mean.

Example 2.3.2. Andrew wants to plan an outdoor party on his birthday. His sister, Karen, has recently returned from Alaska and cannot put up with temperatures above 85° F. His best friend, Bill, does not own a jacket and thus can't stand to be outside when it is below 65° F.

Because he wants both Karen and Bill at the party, Andrew wants to be at least 90% certain that the temperature on his birthday, X, will be within the range such that both will come if it is held outside. If the expected temperature on Andrew's birthday is 75° F and the variance is 25° F, can Andrew plan on hosting his birthday outdoors?

Noting that $\mu = 75, \sigma = \sqrt{25} = 5$, we wish to use Chebyshev's Inequality for the case of $\lambda = 2$ and determine that

$$\begin{aligned}\Pr(|X - 75| \geq 2 \cdot 5) &= \Pr(|X - 75| \geq 10) \\ &\leq \frac{1}{2^2} \\ &= 25\%.\end{aligned}$$

According to this inequality, Andrew can only be 75% certain that both Karen and Bill will show up to an outdoor party, given weather conditions. If this is the best concentration method available, then Andrew should hold his party indoors.

While Chebyshev's Inequality is an important tool, more precise bounds exist for special circumstances. For the special case of random variables known as Poisson trials, the Chernoff Bounds provide exponentially decreasing bounds on the tail distributions.

Definition 2.3.3 (Poisson Trials)**.** Let X be a random variable where $X = \sum_{i=1}^{n} X_i$. If each X_i is a random variable where $\Pr(X_i = 1) = p_i$, $\Pr(X_i = 0) = 1 - p_i$, and all X_i are mutually independent, then each X_i is called a **Poisson trial**.

Definition 2.3.4 (Bernoulli Trials)**. Bernoulli trials** are a special case of Poisson trials where there exists some probability parameter $p_i = p$ for all $i \in \{1, 2, \ldots, n\}$. In other words, all X_i are both mutually independent and identically distributed.

Theorem 2.3.5 (Chernoff Bounds)**.** *Let $X_1, X_2, ..., X_n$ be independent Poisson trials such that $\Pr(X_i = 1) = p_i$. Let $X = \sum_{i=1}^{n} X_i$ and $\mu = \mathrm{E}[X]$. For $0 < \delta < 1$:*

$$\Pr(X \leq (1 - \delta)\mu) \leq \exp\left(-\frac{\mu\delta^2}{2}\right); \tag{2.6}$$

$$\Pr(X \geq (1 + \delta)\mu) \leq \exp\left(-\frac{\mu\delta^2}{3}\right); \tag{2.7}$$

and

$$\Pr(|X - \mu| \geq \delta\mu) \leq 2\exp\left(-\frac{\mu\delta^2}{3}\right). \tag{2.8}$$

Example 2.3.6. The city of Königsberg has taken part in a trial run of a new e-ballot system as a method of raising funds to repair its many bridges. With this system, each of the 200,000 projected votes has an independent probability $p = 0.04$ of being misrecorded in the upcoming mayoral election between two candidates.

The city's electoral board is interested in determining the upper bound on the probability that at least 4.04% of votes will be misrecorded. We let X denote the random variable for the number of misrecorded votes, and let $\mu = 200000 \cdot p = 0.04 \cdot 200000$. Using the Chernoff bound in equation (2.7), we are interested in the case of $\delta = 0.01$:

$$\begin{aligned}\Pr(X \geq (1+\delta)\mu) &= \Pr(X \geq (1+0.01) \cdot 0.04 \cdot 200000) \\ &\leq \exp\left(-\frac{0.04 \cdot 200000 \cdot (0.01)^2}{3}\right) \\ &= \exp\left(-\frac{0.8}{3}\right) \\ &\approx 76\%.\end{aligned}$$

This result gives the board little confidence, only bounding the probability of this unfortunate event as 76%. While this does not make certain that this amount of votes have been misrecorded, lower bounds are desired to reduce the odds of a bad event.

While these examples are illustrative of the use of these tools, they focus on testing the probability of a single "bad" event. Frequently, we are concerned with a number of "bad" events; any of these events could cause the failure of a process or an outcome. In these cases, we are concerned with the probability of *any* bad event occurring. This probability can, in certain instances, be measured precisely using the inclusion-exclusion principle. However, mathematical reasoning in the probabilistic method often only requires bounding the possibility of failure. In such cases, an incredibly effective tool is the Union Bound, also known as Boole's inequality.

Theorem 2.3.7 (Union Bound). *Let $B_1, B_2, ..., B_n$ be a set of events. Then,*

$$\Pr[B_1 \vee B_2 \vee ... \vee B_n] \leq \sum_{i=1}^{n} \Pr[B_i]. \tag{2.9}$$

Example 2.3.8. After a string of electoral reforms, the city of Königsberg has revised its local elections so that candidates must have a majority vote in at least four of the seven precincts. These new precincts all have the same population size. Further, it is projected by the next mayoral election, the city will have a population of 210,000, giving each precinct 30,000 votes. Assuming the probability of a misrecorded vote is unchanged by the next election, what is the upper bound on any precinct having at least 4.4% of the votes misrecorded?

Let X_i denote the random variable for the number of misrecorded votes, and let $\mu_i = 30000 \cdot p = 0.04 \cdot 30000$ for all $i \in \{1, 2, 3, 4, 5, 6, 7\}$. Using the Chernoff bound with the Union bound, we are interested in the case of $\delta = 0.1$:

$$\begin{aligned}\Pr[\bigvee_i (X_i \geq (1+\delta)\mu)] &\leq \sum_i \Pr[X_i \geq (1+0.1)\cdot 0.04 \cdot 30000] \\ &\leq 7 \cdot \exp\left(-\frac{0.04 \cdot 0.1^2 \cdot 30000}{3}\right) \\ &= 7 \cdot \exp\left(-\frac{12}{3}\right) \\ &= \frac{7}{e^4} \\ &\approx 12.82\%\end{aligned}$$

To the relief of the electoral board, there is less than a 13% chance of any precinct having over 4.4% misrecorded. While this is only a bound on our probability, it still provides confidence by providing us with small odds for any bad event.

2.4 Asymptotic Analysis

Often, mathematics concerns itself with the case of explicit formulas. Indeed, there are many cases in which we can enumerate identities or relationships through an explicit formula that yields everything we wish to know, leaving nothing desired. For a function

$$a_n = 3 \cdot 2^n,$$

we are able to calculate a value for all n. However, formulas such as a_n are very rare. An explicit formula may not exist, as is the case of Stirling's Formula for $n!$ or of $\pi(n)$, the number of prime numbers less than or equal to n. Even when we can determine an explicit formula, it may be excessively complicated for calculation, include elements insignificant for the scale of n considered, or both.

To illustrate our notion of insignificant, consider the function

$$b_n = n^3 - n.$$

Depending on the context in which it is used, b_n may be considered as having insignificant elements. Even though it is a relatively simple function to calculate, for large values of n, it possesses more detail than necessary. A mathematician would then use the more compact n^3 in its place. Table 2.1 illustrates this reasoning for several values of n.

While b_n was simple, n^3 clearly suffices for a researcher concerned with values such as $n \geq 100$. In this case, we say that

$$b_n \sim n^3$$

or that b_n is on the order of n^3. For equations much more complicated than b_n, simplification of a formula could even be imperative. We can even extend this

$n =$	n^3	$n^3 - n$	Difference?
10	1000	990	1%
100	1000000	999900	0.01%
1000	1000000000	999999000	0.0001%

Table 2.1: Comparing b_n and n^3

idea further. In many situations, as in this thesis, we may actually be concerned with a class of functions, rather than any one in specific. We may ask which set of functions are asymptotically less than, greater than, or equal to the function order in question. An example of this can be drawn from computer science, where the study of algorithms uses asymptotic analysis extensively.

Suppose that we possess an algorithm to solve a problem. The speed of the algorithm, f, is a function of its input size, n. Because the input to the problem is typically very large, we only concern ourselves with the asymptotic behavior of $f(n)$. What if we were asked for a criteria on which algorithms would be an improvement on our own? If our sole concern was the speed of the algorithm, the obvious response would be any algorithm with a speed function g faster than f.

However, such wordplay leaves us wanting. If our algorithm takes

$$f(n) = n^3$$

hours to compute, we likely don't care for an algorithm taking

$$g(n) = n^3 - n,$$

as we saw in our prior example. The time savings from this improvement would not do much for us, particularly as n grew. Instead, we might state that we are interested in all functions g such that g is dominated by f asymptotically. In other words, we wish that for any $k > 0$, there exists some problem size n_0 such that for all $n > n_0$,

$$|g(n)| \leq k \cdot |f(n)|. \tag{2.10}$$

For $f(n) = n^3$, an example of such an algorithm could be one where

$$g(n) = 100n^2.$$

While $g(n)$ is initially larger than $f(n)$ for $n < 100$, it becomes comparatively more efficient as our problem size n grows. Yet, $100n^2$ is only one such function. A variety of functions for such as

$$g(n) = 15n, e \cdot n^{1.45}, \text{ or } 10^5 \log n$$

would meet our criteria. In place of writing our condition (2.10) each time, we make use of asymptotic notation. With it, (2.10) can be summarized as an interest in all functions

$$g(n) \in o\left(f(n)\right).$$

Table 2.2 displays the family of asymptotic notations and their meanings. This notation will be used extensively in our research when analyzing contagion in large networks.

Table 2.2: The family of asymptotic notations

Notation	Name	Definition
$f(n) \in O(g(n))$	Big O	$\exists k > 0, \exists n_0 : \forall n > n_0,$ $f(n) \leq g(n) \cdot k$
$f(n) \in \Omega(g(n))$	Big Omega	$\exists k > 0, \exists n_0 : \forall n > n_0,$ $f(n) \geq g(n) \cdot k$
$f(n) \in \Theta(g(n))$	Big Theta	$\exists k_1 > 0, \exists k_2 > 0, \exists n_0 : \forall n > n_0,$ $g(n) \cdot k_1 \leq f(n) \leq g(n) \cdot k_2$
$f(n) \in o(g(n))$	Small O	$\forall k > 0 \exists n_0 : \forall n > n_0,$ $\lvert f(n) \rvert \leq k \cdot \lvert g(n) \rvert$
$f(n) \in \omega(g(n))$	Small Omega	$\forall k > 0 \exists n_0 : \forall n > n_0,$ $\lvert f(n) \rvert \geq k \cdot \lvert g(n) \rvert$
$f(n) \sim g(n)$	On the order of	$\forall \epsilon > 0 \exists n_0 : \forall n > n_0,$ $\left\lvert \frac{f(n)}{g(n)} - 1 \right\rvert < \epsilon$

Chapter 3

Inducing Contagion

In this chapter, we discuss previous work on the Linear Threshold Model and introduce the research theme of this thesis.

3.1 Motivation

Our work engages with a broad literature originating in sociology, which studies the diffusion of innovations. While the subject began as an empirical field (e.g., Rogers, 1962), mathematical models were soon introduced by economists and mathematical sociologists (Schelling, 1978; Granovetter, 1978), particularly in game theoretic settings. Parallel to these developments, epidemiologists and mathematicians developed mathematical models of epidemics and probabilistic contagion models which were agnostic in their application. (See Kleinberg, 2007 for a more thorough survey of contagion models).

Within the field of diffusion on social networks, growing recent interest in viral marketing and the availability of data-mining tools have influenced research on algorithmic and economic issues in these models. Our research falls within this scope.

One of the preeminent questions that have arisen is determining the most influential set of nodes, and originates from Domingos & Richardson (2001, 2002). Suppose once again that we are a group of development practitioners seeking to maximize adoption of LLINs within a social network. Given our limited budget, we are able to influence a subset S of individuals to use this new technology. However, once we have done so, we are dependent on their influence to further promote usage of these mosquito nets through word-of-mouth. The question that arises falls within the scope of optimization: how must the set S be chosen to maximize the number of individuals eventually adopting our technology?

This question has been explored in depth by Kempe et al. (2005, 2003). Departing from the work of Domingos & Richardson (2001, 2002), they focus on operational models of social networks, those representing step-by-step dy-

namics of adoption. Using several different operational models, including the Linear Threshold Model, they find that influence functions mapping the set of activated nodes to the number activated by them at time t is submodular, and thus suitable for approximation algorithms. They also prove that influence maximization is NP-hard for the Linear Threshold Model.

In some cases, however, we are interested in seeing how far a given seed can carry its influence. Watts (2000) uses the Linear Threshold Model on arbitrary **random graphs**, graphs where edges are generated by a random process, to explore the probability that a small seed $\Phi(0)$, including the specific case of $|\Phi(0)| = 1$, will not act as a fixed point. As we know from our definition of fixed points in the Linear Threshold model, if a seed is not a fixed point, there will exist some non-seed nodes that will be activated. Using this research, Watts uncovers a condition for global cascades, the case of $\Theta(n)$ nodes being active by convergence, based on a random graph's resulting probability distribution for nodes' degrees.

Finally, we compare the Linear Threshold Model to bootstrap percolation. Bootstrap percolation (Adler, 1991) is a related diffusion model that replicates physical systems such as ferromagnetism. The main distinction is that bootstrap percolation assigns a single threshold r to all nodes (rather than a randomly chosen threshold). At initialization, each node is activated by some probability p independently of all other vertices. For remaining vertices, at each time t, if greater than r neighbors are active, they activate as well. There are several variations on the bootstrap percolation diffusion process. A variety of models including Janson et al. (2012) treat the active state for nodes as a recurrent state. While these models can be useful for modeling physical systems, they hold less value in social networks where thresholds are varied (Valente, 1996).

3.2 Research and Methodology

The questions explored in our research stem from this prior work, inverting the influence maximization question to ask: given an interest in activating some portion of the network, what must the size of our seed be for our process to be successful? Specifically, what is the needed seed size if we desire contagion? In the case of small networks, we revert to the work by Watts (2000) and determine the probability that contagion can be induced by single node or otherwise small seeds. In contrast with similar work on contagion and coordination games by Morris (2000) and Blume (1993, 1995), our research assumes heterogeneity in agent thresholds. Our work falls within the scope of recent research by Acemoglu et al. (2011) and Yildiz et al. (2011) studying the influence of cohesive sets on diffusion within the Linear Threshold Model.

To analyze these questions, our paper borrows from a theoretical concept originating in the field of statistical mechanics, known as mean field theory. Research within the scope of mean field theory attempts to reduce the analytical difficulties of large and complex models of interacting individuals. To do this, the influence of these complex systems on any individual are approx-

imated by a single averaged effect. By approximating the total effect of the system, the graphical structure of our model can be replaced by deterministic equations, greatly easing computation. These approximations have been used extensively in epidemiology, where they have produced surprisingly accurate results in highly connected networks (Gleeson et al., 2012).

The classic mean field behavior involves replacing a complex network structure with the complete graph; this approach is used extensively in game theory under the label of global interaction games (Brock & Durlauf, 2000). Our research borrows this model, studying the Linear Threshold Model as a global interaction game. While we acknowledge other models are useful in our understanding of contagion in social networks, they remain outside the scope of this thesis. A discussion of alternative graphical models in future research is discussed in Chapter 7.

Using our results from small and large networks, we proceed by exploring the impact of a standard feature of real world networks: homophily. Homophily is the tendency of individuals to associate disproportionately with those most similar to themselves, placing themselves as members of specific groups. There is robust evidence of homophily across various demographics (McPherson et al., 2001). However, research modeling the influence of homophily on behavior updating is spare. A recent string of work, led by Matthew Jackson, explores the influence of homophily on a range of issues including consensus times (Jackson & Golub, 2012a), best-response dynamics (Jackson & Golub, 2012b) and the speed of learning (Jackson & Golub, 2009). Our work is most inspired by Jackson & López-Pintado (2012), which explores whether diffusion occurs within Susceptible-Infected-Susceptible models when originating from a small seed. Using a very different set of tools, we pursue a similar line of questioning: given the presence of homophily in our network, how can it influence the minimum size of our seed necessary for contagion?

Our work begins in Chapter 4 by investigating contagion in small networks and determining the efficacy of a seed containing a fixed number of nodes. Most importantly, we introduce a powerful bijection to Dyck Paths which help us to understand whether contagion succeeds or fails. Chapter 5 proceeds by studying large networks. Our technique for these networks generalizes the ideas formed within our study of small networks and applies the tools of the probabilistic method and asymptotic analysis. Through these methods, we derive a lower bound for the size of our seed. Our research ends in Chapter 6, where we derive a new lower bound in homophilous networks. Finally, Chapter 7 discusses potential future avenues of research.

Chapter 4

Contagion in Small Networks

Small networks have two important roles in our understanding of the diffusion of ideas or innovations. First, they are relevant when studying smaller social groups, specifically when they are isolated regarding adoption of the behavior being modeled. Examples of this are the diffusion of innovations within small firms, or of certain ideas or technologies as seen in our introductory example regarding village adoption of LLINs. Each of these examples involve social networks that may be unaffected by actions of other firms or other villages with respect to a certain decision. In these networks, asymptotic analysis is often invalid because of their small size. Instead, we must look at what is actually occurring when watching behavior unfold step by step.

This last point is also of major importance to us. The detailed analysis needed for small networks is an important pedagogical tool for understanding how behavior actually flows through these networks. It reveals behavioral patterns that would not be uncovered through the macro lens used to study large scale networks. These patterns can be of particular importance if we find a bijection between our process with another mathematical object that we already know the properties of, simplifying a great deal of our analysis.

Second, our understanding of contagion stemming from small networks serves as a framework for contagion in large networks. While different methodologies will be introduced, the insights from small networks are directly applied to how we determine whether contagion has occurred.

The first step to understanding whether contagion occurs in a small network is uncovering the existence of cohesive sets.

4.1 Existence of Cohesive Sets

Our research studies the Linear Threshold Model on a complete graph K_n. In a complete graph K_n, we have a simple characterization of cohesive sets.

Lemma 4.1.1. *Consider a complete graph K_n. Let $\mathcal{M}_k$ be the set of vertices with the k highest thresholds. The set $\mathcal{M}_k$ is a cohesive set if and only if*

$$\min_{v_i \in \mathcal{M}_k} \theta_i > \frac{n-k}{n-1}. \tag{4.1}$$

Proof. By Definition 2.2.5, a subset of players $\mathcal{M}$ is a cohesive set if

$$\frac{|\mathcal{M} \cap \mathcal{N}_i(G)|}{|\mathcal{N}_i(G)|} > 1 - \theta_i \text{ for all } v_i \in \mathcal{M}. \tag{2.4}$$

In a simple, complete graph K_n, for any node v_i in the set of vertices $\mathcal{V}$,

$$\mathcal{N}_i(G) = \mathcal{V} \setminus \{v_i\}.$$

Because $|\mathcal{V}| = n$,

$$|\mathcal{N}_i(G)| = n - 1.$$

Similarly, we can state that

$$\mathcal{M} \cap \mathcal{N}_i(G) = \mathcal{M} \cap (\mathcal{V} \setminus \{v_i\}).$$

Because of the associative law of sets and that $\mathcal{M} \subseteq \mathcal{V}$,

$$\begin{aligned} \mathcal{M} \cap \mathcal{N}_i(G) &= \mathcal{M} \cap (\mathcal{V} \setminus \{v_i\}) \\ &= (\mathcal{M} \cap \mathcal{V}) \setminus \{v_i\} \\ &= \mathcal{M} \setminus \{v_i\}. \end{aligned}$$

Therefore,

$$\begin{aligned} |\mathcal{M} \cap \mathcal{N}_i(G)| &= |\mathcal{M} \setminus \{v_i\}| \\ &= |\mathcal{M}| - 1. \end{aligned}$$

Replacing these values into equation (2.4) results in the condition:

$$\frac{|\mathcal{M}| - 1}{n-1} > 1 - \theta_i \text{ for all } v_i \in \mathcal{M} \tag{4.2}$$

which is equal to

$$\frac{n - |\mathcal{M}|}{n-1} < \theta_i \text{ for all } v_i \in \mathcal{M}. \tag{4.3}$$

If the set $\mathcal{M}$ is a set of size k, denoted as $\mathcal{M}_k$, this condition is rewritten as

$$\theta_i > \frac{n-k}{n-1} \text{ for all } v_i \in \mathcal{M}$$

and equivalent to (4.1):

$$\min_{v_i \in \mathcal{M}_k} \theta_i > \frac{n-k}{n-1}. \tag{4.1}$$

□

For a complete graph K_n, this means that the existence of a cohesive k-set requires the existence of at least k nodes with thresholds greater than

$$\frac{n - |\mathcal{M}_k|}{n-1} = \frac{n-k}{n-1}.$$

Intuitively, the threshold of each of the k nodes must be greater than the fraction of the graph, excluding themselves, that is not in the k-set.

To reinforce the implications of the cohesive set, we also redefine the activation process, equation (2.2), in the context of the complete graph.

Lemma 4.1.2. *For a complete graph K_n, given a set $\Phi(t)$ of active nodes at time t, a node v_i is active at time $t+1$ if:*

$$|\Phi(t) \setminus \{v_i\}| \geq \lceil \theta_i \cdot (n-1) \rceil \Rightarrow v_i \in \Phi(t+1). \tag{4.4}$$

Proof. This proof follows a similar argument as Lemma 4.1.1. Because

$$\mathcal{N}_i(G) = \mathcal{V} \setminus \{v_i\},$$

a node v_i is active at time $t+1$ if:

$$|\Phi(t) \setminus \{v_i\}| \geq \kappa_i \Rightarrow v_i \in \Phi(t+1). \tag{4.5}$$

Next, we note that $\kappa_i = \lceil \theta_i \cdot d_i \rceil$. Because we are dealing with the complete graph, $d_i = n-1$ for all nodes v_i. Thus,

$$\kappa_i = \lceil \theta_i \cdot (n-1) \rceil$$

so (4.5) is rewritten as (4.4):

$$|\Phi(t) \setminus \{v_i\}| \geq \lceil \theta_i \cdot (n-1) \rceil \Rightarrow v_i \in \Phi(t+1). \tag{4.4}$$

□

Combining these two lemmas, a member of a cohesive set $\mathcal{M}$ can only be activated at time $t+1$ if another member of $\mathcal{M}$ is already activated by time t. Using backward induction, this means that a member of the cohesive set can only be activated if some member of $\mathcal{M}$ is in the seed $\Phi(0)$. For this reason, we begin by examining when cohesive sets of a certain size exist. This leads to an understanding how large each seed must be and which nodes it must include.

This analysis begins with our first use of the ideas of mean field theory. Because all nodes are connected in the complete graph, we can treat these nodes as indistinguishable with the exception of their thresholds. We take advantage of this by abstracting from our graph structure. Instead, we use the threshold values to represent our network using a balls and bins model. This model studies problems as a case of r balls being thrown into one of s bins independently and uniformly at random. In the Linear Threshold Model on a complete graph K_n, the balls represent each vertices $\mathcal{V}$. Each ball i is placed into a bin according to

the value of $\theta_i \cdot d_i$. The bins are a division of the threshold distribution interval (0, n - 1]. Throughout the use of this model, balls will be used interchangeably with other terms previously used to denote nodes in our diffusion model.

In the case of our small network, we study a division of the interval into bins of size $1/(n-1)$. For simplicity, we then scale our analysis such that balls represent each $\kappa_i = \lceil (n-1) \cdot \theta_i \rceil$ and bins are of unit size 1 in the interval $(0, n-1]$. Each ball is placed in the bin labeled κ_i. An example of this threshold division is shown in Figure 4.1.

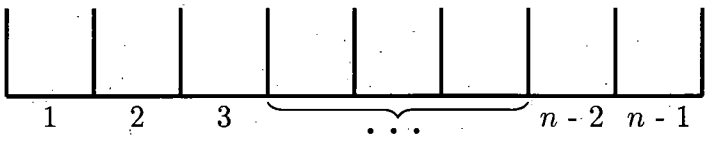


Figure 4.1: The bin distribution for K_n

Within this model, a cohesive set of size k corresponds to having k nodes placed in the $k-1$ highest valued bins. The converse of the statement is that a cohesive set of size k or smaller does not exist when there are fewer than j nodes in the $j-1$ highest valued bins for all $2 \leq j \leq k$. Because our thresholds are distributed uniformly at random, we are interested in the *probability* that there is no cohesive set of size k or smaller in our network K_n. Therefore, we think about **distribution classes**, an abstraction of the threshold distribution. In these classes, we care (i) which bin a node's threshold falls into, and (ii) the linear order of the thresholds within the bin. In addition, we will make great use of the following bijection.

Lemma 4.1.3. *Let D_n be the set of all non-decreasing sequences of positive integers of size k, or sequences of integers $1 \leq a_1 \leq \ldots \leq a_k$ where $a_i \leq i$. Then, $|D_k| = C_k$, the k^{th} Catalan number.*

The proof of this is lemma is the solution to an exercise in Stanley (2001), and is commonly available, including at Stanley (n.d.). It provides a bijection from the sequence to a known Catalan problem. One such problem is that of **Dyck Paths**, or "mountain ranges," formed with k upstrokes and k downstrokes that all stay above the original line. Table 4.1 displays the list of possibilities of mountain ranges for $0 \leq k \leq 3$. The number of possibilities for each k is equal to

$$C_k = \frac{1}{k+1}\binom{2k}{k}.$$

k	Possible Ranges	C_k
0	$\emptyset$	1
1		1
2		2
3		5

Table 4.1: Mountain Ranges

With this lemma in hand, we are able to count the number of possible cohesive sets for a given distribution of nodes in a complete graph. Choosing a random threshold θ_i results in throwing a ball into a random bin. We want to know how many distributions exist where there are no cohesive sets of size k or smaller. The following lemma is the first step towards this measurement.

Lemma 4.1.4. *Let B_{k+1} be the set of distributions of $k+1$ balls into k bins where there are at most j balls in the j highest valued bins, for all $1 \leq j \leq k-1$. Then, $|B_{k+1}| = C_k$.*

Proof. Our proof will rely on a bijection with the set D_k from Lemma 4.1.3. Consider a sequence of non-decreasing positive integers in D_k. Further, nodes are presented as ordered by threshold so that v_1 is the node with the lowest value θ_i and v_{k+1} is the node with the highest threshold θ_i in our network. Let each a_i indicate the bin in which node v_{i+1} is placed, thus skipping the node v_1. Given $a_i \leq i$, for any $1 \leq j \leq k$, there are at most j balls (or nodes) in j highest valued bins, those with labels in the set $\{k-j+1, \ldots, k\}$. This is equivalent to our restrictions for the set B_n.

Node v_1 can be excluded from the mapping. To understand this, note that $a_1 = 1$ and a_1 indicates the placement of the node with the second lowest threshold, v_2. Since v_1 has a lower threshold, and 1 is the lowest possible bin, node v_1 must be in bin 1 in each case. Including it in our mapping is therefore unnecessary.

With all nodes now considered, we describe our bijection from B_{k+1} to D_k. Because we know from Lemma 4.1.3 that $|D_k| = C_k$, we also have $|B_{k+1}| = C_k$. This two step bijection is illustrated in Figure 4.2. For $i > 1$, each listing of κ_i values in ascending order is immediately a bijection to D_k. The mountain ranges illustrate the Catalan numbers. Because we do not map v_1 in our bijection and it is consistent across each case, we ignore it in our visual bijection.

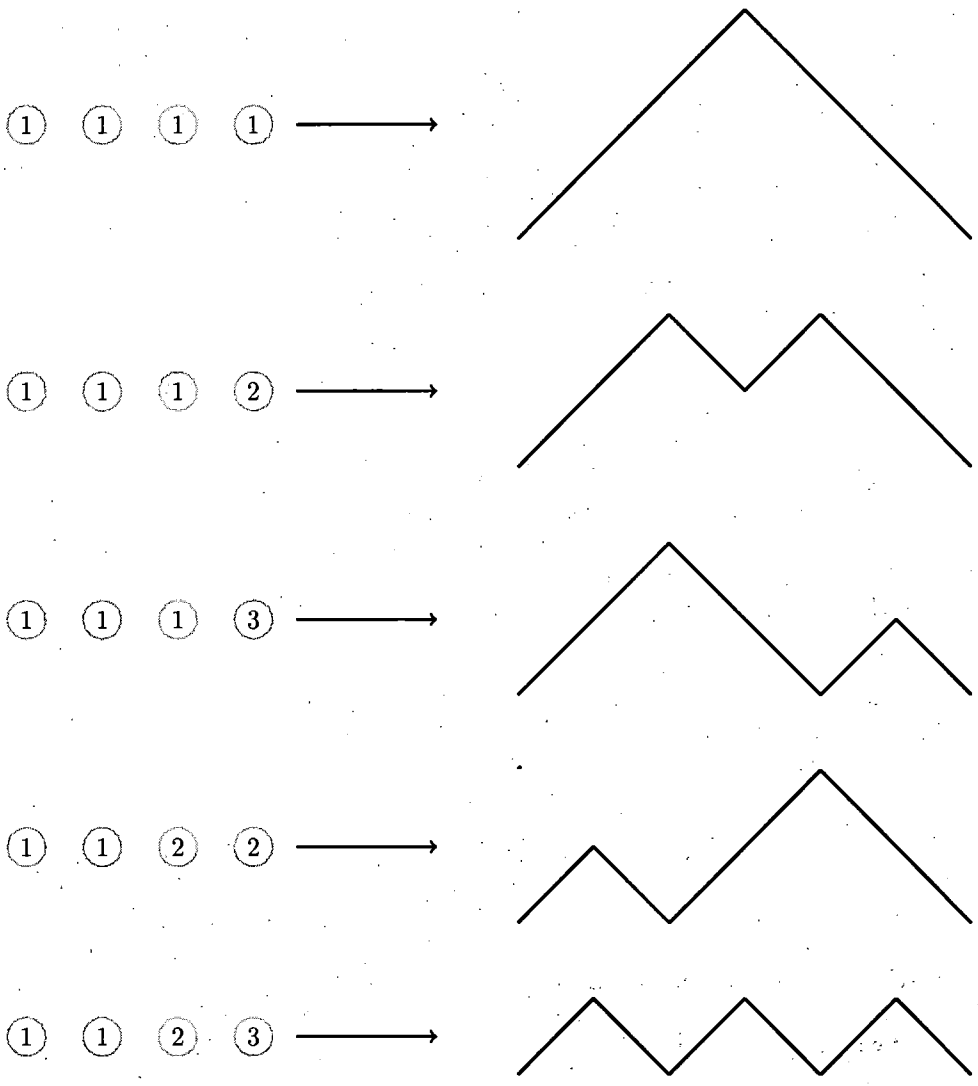


Figure 4.2: A visual example of the bijection from B_4 to M_3, mountain ranges of length 6.

Visually, we state our bijection as the following. We read our sequence from left to right. Each time an a_i is encountered, there is a single upstroke in our mountain. Whenever there is an ascent, meaning $a_i < a_{i+1}$, the mountain range has $a_{i+1} - a_i$ downward strokes. This downward stroke occurs in between the upward strokes of each of the two integers.

□

Lemma 4.1.5. *The number of threshold distribution classes without a non-trivial cohesive set for K_{n+1} is counted by the nth Catalan number, C_n.*

Proof. For K_{n+1} to not have any non-trivial cohesive sets, there must be at most than j nodes in the j highest valued bins, for all $1 \leq j \leq n$. Thus, each valid distribution of thresholds is also a valid distributions of balls in bins under B_n. (This, appropriately, means that the first bin, with label 1, has no restriction on the number of nodes it holds. Nodes in the first bin can only contribute to the trivial cohesive set of $\mathcal{M} = \mathcal{V}$.) The key assumption here is again that the node with the lowest valued threshold, v_1, must always be in the first bin, because v_2 must also be placed in bin 1 and $\theta_2 > \theta_1$ by construction. v_1 can therefore always be ignored, as it is the same in each mapping.

Because of these matched restrictions, we can state that the number of threshold distribution classes without a non-trivial cohesive set for K_{n+1} is counted by B_{n+1}. By Lemma 4.1.4, $|B_{n+1}| = C_n$, the nth Catalan number. □

This lemma treats nodes as indistinguishable. However, to calculate the odds of contagion, we must consider the case where our nodes are distinguished. To do this, we use **labeled distribution classes**.

Corollary 4.1.6. *Assume our social network is represented by the graph K_{n+1}. Let the size of the set of distribution classes we are interested in be denoted by $R(n)$. Then, the number of labeled distribution classes of interest is*

$$(n+1)! \cdot R(n).$$

Proof. Recall that thresholds are drawn uniformly at random from the interval $(0, 1]$. Therefore,

$$\Pr(\theta_i = \theta_j) = 0 \text{ for any } v_i, v_j \in \mathcal{V}.$$

This means that there is a linear ordering to our thresholds. The linear ordering of $n+1$ objects is counted by $(n+1)!$. Thus, for each of our $R(n)$ distribution classes of interest, there are $(n+1)!$ ways to label the nodes $\{v_1, \ldots, v_{n+1}\}$. □

Their are two immediate consequences of this corollary. First, the total number of labeled distribution classes is

$$(n+1)! \cdot \binom{n + (n+1) - 1}{n+1} = (n+1)! \cdot \binom{2n}{n+1}.$$

Second, we can enumerate the number of labeled distribution classes without cohesive sets of size k or smaller. To do this, we first distribute the k nodes with the highest thresholds so that there are no more than j nodes in the j highest bins for $1 \leq j \leq k$. This is enumerated by C_n, as shown previously. For the remaining $n+1-k$ nodes, they can be distributed in any number of ways we can place $n+1-k$ undistinguished balls in $n-k$ boxes. This is enumerated by

$$\binom{(n+1-k) + (n-k) - 1}{n-k} = \binom{2(n-k)}{n-k}.$$

Therefore, the number of labeled distribution classes without cohesive sets of size k or smaller is

$$(n+1)! \cdot \binom{2(n-k)}{n-k} \cdot C_k. \tag{4.6}$$

Example 4.1.7. Let our social network be represented by the graph K_7. Our thresholds θ are distributed independently and uniformly at random. We first treat all nodes as indistinguishable with the exception of their thresholds. What is the probability that there are no cohesive sets of size 3 or smaller?

Using Lemma 4.1.5, we know that for K_4, there are C_3 distribution classes that do not contain non-trivial cohesive sets. Next, we consider K_4 as a subgraph of K_7. Specifically, we map K_4 onto K_7 by increasing κ_i for all nodes in K_4 by 3, with the exception of $v_i \in \mathcal{V}(K_4)$ for whom $\theta_i = \min_{v_j \in \mathcal{V}(K_4)} \theta_j$. Now each individual in the subgraph K_4 is in one of the 3 highest bins as they were originally, except the individual with the lowest threshold. An example mapping, with thresholds for $\mathcal{V}(K_7) \backslash \mathcal{V}(K_4)$ undetermined, is displayed in Figure 4.3.

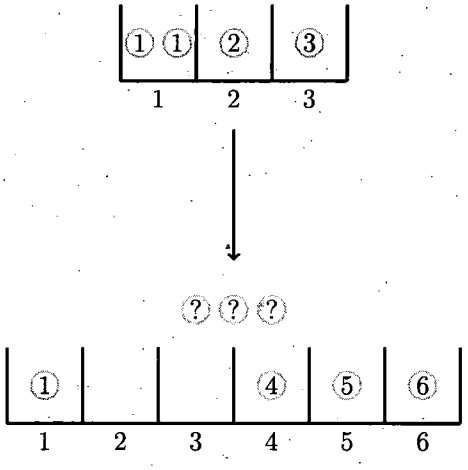


Figure 4.3: Mapping K_4 into K_7. There are 3 unknown thresholds.

To avoid the existence of a cohesive set, we cannot have more than 3 nodes in the 3 highest bins. Therefore, the remaining nodes must have a threshold such that $\kappa_i \in \{1, 2, 3\}$ for all $i \notin \mathcal{V}(K_4)$. We thus have three nodes that we can place in any of the three bins, which can be done in $\binom{3+3-1}{3} = 10$ ways.

Finally, we can calculate the probability that our network has no cohesive sets of size 3 or smaller. By Corollary 4.1.6, we can say that there are

$$7! \cdot \binom{12}{7} = 3,991,680$$

total labeled distribution classes of balls within bins. There are

$$7! \cdot \binom{6}{3} \cdot C_3 = 504,000$$

labeled distribution classes with no cohesive sets of size 3 or smaller. Therefore, the probability that there will be no cohesive sets of size 3 or smaller is

$$504000/3991680 \approx 13\%.$$

4.2 Contagion when $|\Phi(0)| = 1$

The question set forth in Example 4.1.7 leads to our first major theorem.

Theorem 4.2.1. *Assume a network K_{n+1}. If there are no non-trivial cohesive sets, then a seed consisting of any individual node can instigate contagion.*

Proof. Without non-trivial cohesive sets, the restriction of a cohesive set only being activated if a member of the set is in the seed $\Phi(0)$ only applies to the set of all vertices. We then know that for any set of nodes $\mathcal{M}$,

$$\frac{(n+1) - |\mathcal{M}|}{(n+1) - 1} \geq \theta_i \quad \text{for some} \quad v_i \in \mathcal{M}.$$

Using this, we prove diffusion will occur to the entire network by induction.

Starting with any individual node s as our seed, it follows that $|\Phi(0)| = 1$. Then, for the set of nodes $\mathcal{M}_n = \mathcal{V}(K_{n+1}) \setminus \{s\}$,

$$\frac{n+1-|\mathcal{M}_n|}{n+1-1} = \frac{1}{n} \geq \theta_i \quad \text{for some} \quad v_i \in \mathcal{M}_n$$

and so

$$1 \geq \theta_i \cdot n \quad \text{for some} \quad v_i \in \mathcal{M}_n.$$

This last inequality states that some $v_i \in \mathcal{M}_n$ is now activated as a result of our seed $\Phi(0)$ because of its sufficiently low threshold

$$\theta_i \cdot n \leq |\Phi(0)|.$$

We continue repeating this process. At every time t, $\mathcal{M}$ is considered to be the set $\mathcal{V}(K_{n+1}) \setminus \Phi(t)$. Therefore, at every time t, if $\mathcal{M} \neq \mathcal{V}(K_{n+1})$ then

$$\frac{(n+1) - ((n+1) - |\Phi(t)|)}{(n+1) - 1} \geq \theta_i \quad \text{for some} \quad v_i \in \mathcal{M}.$$

Because $\Phi(t)$ is always an integer, we again ignore the ceiling function and note that our activation requirement is met:

$$\Phi(t) \geq \theta_i \cdot n \quad \text{for some} \quad v_i \in \mathcal{M}.$$

Therefore, for any set of unactivated nodes, there will always be at least one node that is activated in the next time period. Due to the finite number of nodes, every node will eventually become active and thus the behavior is contagious. □

Corollary 4.2.2. *Assume a network represented by K_{n+1}. Then,*

$$\Pr[\text{ Contagion } | \ |\Phi(0)| = 1] = \frac{C_n}{\binom{2n}{n+1}} = \frac{1}{n}.$$

Intuitively, the probability that any singleton seed can induce contagion is $\frac{1}{n}$.

Proof. We know from Theorem 4.2.1 that diffusion will occur from any seed of size one when there are no non-trivial cohesive sets. From Lemma 4.1.5 and Corollary 4.1.6, we know that there are $(n+1)! \cdot C_n$ labeled distribution classes in which this occurs. Further, we know that the number of total labeled distribution classes is

$$(n+1)! \cdot \binom{(n+1)+n-1}{n+1} = (n+1)! \binom{2n}{n+1}.$$

Therefore,

$$\begin{aligned}
\Pr[\text{ Contagion } \mid |\Phi(0)| = 1] &= \frac{(n+1)! \cdot C_n}{(n+1)! \cdot \binom{2n}{n+1}} \\
&= \frac{(n+1)!}{n+1}\binom{2n}{n} \cdot \frac{1}{(n+1)! \cdot \binom{2n}{n+1}} \\
&= \frac{1}{n+1} \cdot \frac{(2n)!}{n!n!} \cdot \frac{(n+1)!(n-1)!}{(2n)!} \\
&= \frac{1}{(n+1)!n!} \cdot (n+1)! \cdot (n-1)! \\
&= \frac{1}{n}.
\end{aligned}$$

□

This result is of particular importance when we cannot directly observe the threshold of any particular node in a network. In that case, and if we are limited to a seed of size one, we now know the probability of contagion from any singleton seed, including the node with the lowest threshold. However, we see that as n increases, the probability of there not being a non-trivial cohesive set quickly diminishes:

$$\Pr[\text{ Contagion } \mid |\Phi(0)| = 1] = \frac{1}{n} \to 0 \quad \text{as } n \to \infty.$$

We now consider a new situation: what if we can observe which individual has the highest threshold in our network and choose her as our seed? How much would this observation improve our odds of contagion? To answer these questions, we first introduce a formal definition of the Dyck Paths. Next, we introduce a important modification of these paths that we will use in our main result.

Definition 4.2.3 (Dyck Paths)**.** A Dyck Path is a sequence of unit movements up (U) or down (D) constructing a walk where no initial segment of the string has more Ds than Us. The number of Dyck paths for any given value of n is counted by C_n.

Lemma 4.2.4. *The number of Dyck $(n+1)$-paths ending with DD is enumerated by*

$$\frac{3(2n)!}{(n+2)!(n-1)!}.$$

Proof. These paths are enumerated by the **Lobb numbers** $L_{n,1}$, for which this is the equation. A generalized proof of this fact is provided in Section 4.3. □

With our lemmas at hand, we can finally count the distribution classes where there exists a seed that can induce contagion.

Theorem 4.2.5. *Let our social network be represented by K_{n+1}. The number of distribution classes where there exists a singleton seed that can induce contagion is enumerated by*

$$\frac{3(2n)!}{(n+2)!(n-1)!}.$$

Proof. A cohesive set can only become active if a member of the set is already active. Also, the set of all cohesive sets is nested, such that the most restrictive set lies within all other cohesive sets. This most restrictive set is, by definition, the cohesive set that includes the nodes with the highest thresholds. Therefore the node with the highest threshold is a member of each cohesive set.

Our process begins by preselecting this node with the highest threshold as our seed of size one. We note that there are distribution classes that do not require the highest threshold node as the seed in order for a behavior to be contagious. However, these distribution classes will nonetheless witness contagion when the highest seed is selected and are thus considered.

We will describe a bijection from the n non-seed nodes to the 1st fold self-convolution of the Catalan numbers. We begin our process in a manner similar to Lemma 4.1.4, our previous bijection to Dyck paths. We read the ordered and reindexed list of nodes from left to right, ascenting by one (recording a U) each time we read a node i and descenting by $k_{i+1} - k_i$ (recording $k_{i+1} - k_i$ many Ds) between any nodes i and $i+1$.

Here, we make two important modifications. First, we no longer ignore any nodes. All n nodes are considered and read. As before, there are at least as many Us as Ds in any initial string, otherwise activation of the following node could not be ensured. Second, we must take care of the fact that we now have one more node than we do ascents. To do this, we record a final D at the end of our string to signify the end our process.

Because there was also a D immediately prior, to denote the completion of the last bin, each path ends in DD. An example of this mapping is provided in Figure 4.4.

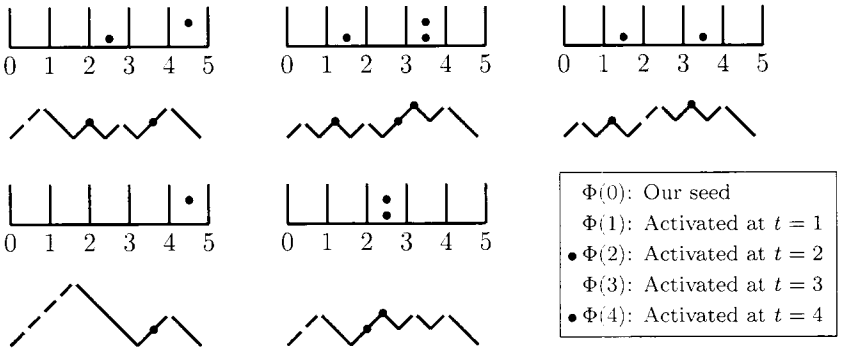


Figure 4.4: Sample mappings from contagion in K_6 to Dyck paths ending in DD.

The Dyck $(n+1)$-paths ending in DD are enumerated by

$$\frac{3(2n)!}{(n+2)!(n-1)!}$$

by Lemma 4.2.4, and thus we achieved our desired enumeration. □

Corollary 4.2.6. *Assume a network represented by K_{n+1}. Then,*

$$\Pr[\text{ Contagion } \mid \Phi(0) = \{v_i \mid \theta_i = \max_{v_j \in \mathcal{V}} \theta_j\}] = \frac{3(2n)!}{(n+2)!(n-1)!} \cdot \frac{1}{\binom{2n}{n+1}} = \frac{3}{n+2}.$$

Proof. This proof follows the structure of Corollary 4.2.2. Our denominator is modified for the case of K_{n+1}. The numerator arises from the number of possible distribution classes, as determined by Theorem 4.2.5. Finally, we complete the requisite algebra:

$$\begin{aligned}\Pr[\text{ Contagion } \mid \Phi(0) = \{v_i \mid \theta_i = \max_{v_j \in \mathcal{V}} \theta_j\}] &= \frac{3(2n)!(n+1)!}{(n+2)!(n-1)!} \cdot \frac{1}{(n+1)! \cdot \binom{2n}{n+1}} \\ &= \frac{3(2n)!}{(n+2)!(n-1)!} \cdot \frac{(n+1)!(n-1)!}{(2n)!} \\ &= \frac{3}{n+2}.\end{aligned}$$

□

Diffusion is more likely to occur in a network K_n, for $n > 3$, when we can designate the node with the highest threshold as our seed. However, we note that whether or not the threshold of our seed can be controlled, the probability of contagion remains at $\Theta(1/n)$. Therefore, our probability of contagion quickly approaches zero as a network grows. To further improve our odds, we examine the case of picking a seed consisting of multiple nodes.

4.3 Contagion when $|\Phi(0)| = k$

We now consider the case where our seed can be the size of any fixed integer k. By increasing our seed, we can significantly increase our odds. To do so, we must first return to a previously mentioned idea, the Lobb numbers $L_{n,k}$

Definition 4.3.1 (Lobb numbers). The Lobb number $L_{n,k}$, where $0 \leq k \leq n$, counts the number of arrangements of $n+k$ positive ones and $n-k$ negative ones such that every partial sum is nonnegative.

Lemma 4.3.2 (Lobb (1999)). *The Lobb numbers are computer by*

$$L_{n,k} = \frac{2k+1}{n+k+1}\binom{2n}{n+k}.$$

It follows that $L_{n,0} = C_n$, $L_{n,n} = 1$, and $L_{n,n-1} = 2n-1$.

Theorem 4.3.3.

$$\sum_{k=1}^{n} L_{n,k} = \binom{2n}{n+1}$$

which is the number of ways to place $n+1$ balls in n bins.

Proof. The nth row sum of the Lobb numbers is known to be the nth central binomial coefficient (Koshy, 2009):

$$\sum_{k=0}^{n} L_{n,k} = \binom{2n}{n}.$$

Further, we know that $L_{n,0} = C_n = \frac{1}{n+1}\binom{2n}{n}$. Therefore,

$$\begin{aligned}
\sum_{k=1}^{n} L_{n,k} &= \binom{2n}{n} - \frac{1}{n+1}\binom{2n}{n} \\
&= \frac{n+1}{n+1}\binom{2n}{n} - \frac{1}{n+1}\binom{2n}{n} \\
&= \frac{n}{n+1} \cdot \frac{(2n)!}{n!n!} \\
&= \frac{(2n)!}{(n+1)!(n-1)!} \\
&= \binom{2n}{n+1}
\end{aligned}$$

as desired. □

Lemma 4.3.4. *There exists a bijection between the sequence of $+1$s and -1s counted by the Lobb numbers $L_{n,k}$ and Dyck paths with $n+k$ ascents U and $n-k$ descents D, where there are more entries U than D in any initial string.*

Proof. This proof is analogous to the Catalan equivalency proof for these objects when $k = 0$. Each $+1$ is equivalent to a D and each -1 is equivalent to a U. □

A more in-depth introduction to the Lobb numbers is provided in Koshy (2009). We now generalize an earlier result from the previous section, showing a relationship between contagion induced from a seed of size k to the Lobb number $L_{n,k}$.

Theorem 4.3.5. *Let our social network be represented by K_{n+1}. The number of distribution classes in which a seed consisting of the k nodes with the highest thresholds will induce contagion, but a seed consisting of the j nodes with the highest thresholds for $j < k$ could not, is enumerated by $L_{n,k}$. The case of $k = 0$ represents the cases where any singleton seed can induce contagion.*

Proof. We read the list of $n + 1$ nodes in our network from those with the smallest to largest values of κ_i and reindex them appropriately. Each time a node is read, we write a U in our Dyck path string. A number of descents, D, equal to the ascent in values κ_i are recorded between nodes. Thus, if $\kappa_1 = 1$ and $\kappa_2 = 3$, the string from κ_1 to κ_2 would read $UDDU$. Finally, a last D is recorded to mark the end of the last bin, giving us our nth entry D.

We begin with the special case of $k = 0$. If $k = 0$, we will choose to ignore κ_1, representing the node with the lowest threshold, in each string. If there are no non-trivial cohesive sets, we have shown that the node with κ_2 is always in the first bin, and so we will always begin with a U. Thus, we do not yet need to concern ourselves with the case that the lowest threshold node. Therefore, we are counting the number of Dyck paths of length n with at least as many Us in Ds in any initial string. As shown in Lemma 4.1.5, this corresponds to the case of no non-trivial cohesive sets, where any singleton seed can induce contagion. These are counted by $C_n = L_{n,0}$. This special case is complete.

There is another special case for $k = 1$. The case $k = 1$ includes threshold distributions valid for $k = 0$. Although those distribution classes do not require the highest threshold node to act as the seed, they still require a single seed to be activated and so they are counted. This fact is consistent with Theorem 4.3.3, which, with this theorem, states that the sum of distribution classes for which a seed consisting of the k nodes with the highest thresholds is a necessary and sufficient condition, over all $1 \leq k \leq n$ is equal to the total number of distribution classes of nodes in bins.

The mapping for $k = 1$ proceeds similarly to the case $k = 0$. However, we never remove the first node. We know that the first node must still always be in the first bin, because the next bin requires 2 active nodes for any nodes in it to activate. Therefore, a singleton seed in the topmost bin must be coupled with a second node in the first bin that is activated by it. We record our Ds and Us as we did previously, but do not mark the end of our path with a final D as we did previously. This gives us a mapping of $n + k = n + 1$ entries U and $n - k = n - 1$ entries D. This path is then enumerated by $L_{n,1}$ based on Lemma 4.3.4.

If $k > 1$, we slightly modify our bijection. Note that size of the seed excluding the member with the highest threshold is $k - 1$. We start with our path of entries U from lowest to highest threshold nodes with descents D in between representing moving to the next bin. There are several descents along the way that prevent any initial string from having at least as many Us as Ds. However, we have yet to take into account the influence of our seed.

In this current path, we have $n - 1$ entries D. We will remove $k - 1$ of these entries. As we read the path from left to right, we want to delete the the first $k-1$ descents because they are the ones we have compensated for with our seed. To do this, we remove D entries from left to right until $k - 1$ entries are gone. At each point where there was a continuous substring of descents D, we note the number $j < k - 1$ of entries D removed from this substring, and place j entries U starting at the index where we began removing them, indicating that a node in the seed was activated in place of the descent. Figure 4.5 provides an example of one such mapping.

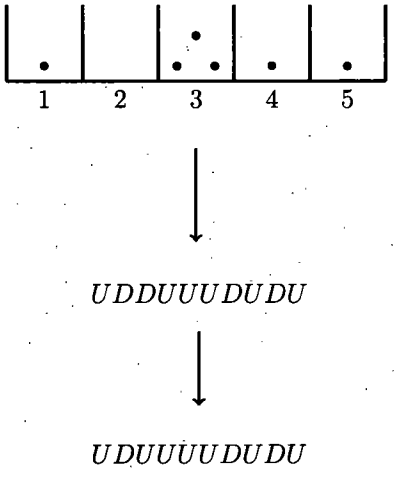


Figure 4.5: A distribution class for K_5 where $k = 2$.

To prove this path is equivalent to the Dyck paths denoted by $L_{n,k}$, we prove equivalency to each requirement separately. First, note that we have removed $k-1$ of the $n-1$ total entries D. Therefore, our path now has $n-1-(k-1) = n-k$ entries D. Similarly, we have added $k-1$ entries U to the existing amount, $n+1$. Now, our total number of entries U is $n + 1 + k - 1 = n + k$. Note that these are unique to k, and so cases needing $j < k$ nodes in the seed are not counted. Therefore, the number of entries meet the first requirement for our Dyck paths.

Finally, we must prove that in any initial string, there are at least as many Us as Ds. However, this must be the case, or contagion could not be induced on these networks. For any first $k - 1$ entries of D we encounter, there must have been at least $k-1$ entries U prior to it. This is because of our replacement of $k - 1$ prior descents with ascents.

Let m denote the number of U entries recorded not through this replacement process. If we encounter a jth entry D, where $j > k-1+m$, and we still only have $k-1+m$ ascents recorded, this means contagion has not occurred. It means we have reached the bin labeled $j > k-1+m$ which cannot be activated unless at least j nodes are activated, equivalent to j entries U having been recorded. Therefore, we meet the requirement that there are at least as many ascents as descents in any given initial string. We have completed our mapping to the Dyck paths enumerated by the Lobb numbers according to Lemma 4.3.4. □

Corollary 4.3.6. *Let our social network be represented by K_{n+1}. Let T_k denote the event our seed consists of the nodes with the k highest thresholds. Then,*

$$\Pr[\text{ Contagion } | \ T_k] = \frac{\sum_{i=1}^{k} L_{n,i}}{\binom{2n}{n+1}}$$

Proof. This result follows directly from Theorem 4.3.5. □

Chapter 5

Contagion in Large Networks

While small networks have several applications and serve as important pedagogical tools, we now focus our attention on large networks. In our research, we do not classify large networks by some threshold size X, stating that all networks G where $|\mathcal{V}(G)| \geq X$ are classified as large. Large networks are, by this criteria, a misnomer. Instead, we investigate properties of graphs as $|\mathcal{V}(G)| \to \infty$. Combined with our use of asymptotic analysis, readers can themselves calculate the bounds on the errors of our calculations for whichever network size they are interested in.

Aymptotic analysis presents many benefits for analyzing social networks. However, several modifications to the Linear Threshold Model will simplify our analysis. These modifications will be used from this point forward.

5.1 A Linear Threshold Model for Asymptotic Analysis

Let our social network be represented by a complete graph K_n for some large value n. Each node v_i in K_n has a threshold drawn independently and uniformly at random from the interval $[0, n)$. Here, we have made three very important modifications.

1. θ_i is now defined in a manner similar, but not equivalent, to the definition of κ_i in the original model. This notion of threshold avoids the awkward, and unnecessary, use of fractions to describe our thresholds. As we will see, we have no use for a ceiling function of our thresholds times the size of our network, denoted by κ_i, in this model.

2. We allow $\theta_i = 0$ for any $v_i \in \mathcal{V}(K_n)$. Because $\Pr[\theta_i = 0] = 0$ for any node v_i, this modification is inconsequential to our results

3. As with the second assumption, we have modified the upper bound of our threshold interval somewhat inappropriately. This upper bound, consisting of the interval $(n-1, n)$ is, to some extent, illogical. It directly translates to needing more individuals than available to be activated as our threshold. Still, such a modification provides a cleaner exposition in our analysis. Further, its consequence is less than $1/n$, which becomes inconsequential as $n \to \infty$.

With these modifications, we can evaluate the concentration of our bins. In contrast with the small networks of Chapter 4, we do not need to concern ourselves with all possible threshold distributions. Because of the Law of Large Numbers, we focus on bounding the deviation of our realized threshold distribution from its expected profile.

5.2 Concentrating Bins

Let $0 < \alpha < 1$ be a constant to be characterized later. We partition $[0, n)$ into

$$M = n^{1-\alpha}$$

intervals, each length n^α. For $1 \leq i \leq M$, the ith interval is

$$J_i = \left[(i-1) \cdot n^\alpha, i \cdot n^\alpha\right).$$

Let B_i denote the ith *bin*, which is the set of nodes whose thresholds are in J_i. Let $\mu_i = \mathrm{E}[|B_i|]$. A bin is **concentrated** if $|B_i| = \mu_i \pm o(\mu_i)$ *with high probability* (*whp*) as n tends to infinity.

Lemma 5.2.1. *Let K_n represent our social network of n individuals as $n \to \infty$. Simultaneously, each bin B_i contains $n^\alpha \pm \sqrt{3n^\alpha \log n}$ nodes* whp.

Proof. For $1 \leq i \leq M$, let X_i be the random variable indicating the number of nodes in bin B_i. These variables are identical (but not independent) random variables that correspond to n Bernouli trials with success probablilty $1/M$. For each X_i,

$$\mathrm{E}[X_i] = \mu = \frac{n}{M} = n^\alpha.$$

To achieve our desired concentration, our deviation from the mean, $|X_i - \mu|$ is bounded above by $o(n^\alpha)$. Consider the placement of a node into bin B_i as a Poisson trial with success probability $1/M$. Each X_i is the sum of n Poisson trials. Therefore, we can use the Chernoff bound on each of these variables. We have

$$\Pr\left[\bigvee_{1 \leq i \leq M} \left(|X_i - \mu| \geq \delta\mu\right)\right] \leq \sum_{1 \leq i \leq M} \Pr\left[|X_i - \mu| \geq \delta\mu\right] \leq 2M \cdot e^{-\mu\delta^2/3}.$$

The first inequality follows from a union bound (2.9). The second inequality follows from the lower Chernoff bound (2.6). To obtain simultaneous concentration for all bins *whp*, we need

$$2M \cdot e^{-\mu\delta^2/3} \to 0 \text{ as } n \to \infty.$$

Let $\delta = \sqrt{\frac{3\log n}{n^\alpha}}$. We have

$$\begin{aligned} 2M \cdot e^{-\mu\delta^2/3} &= 2M \cdot \exp\left(-\frac{\mu\,\delta^2}{3}\right) \\ &= 2\exp\left((1-\alpha)\log n - \frac{n^\alpha\delta^2}{3}\right) \\ &= 2\exp\left(-\alpha\log n\right) \\ &= 2n^{-\alpha} \to 0. \end{aligned}$$

Finally, we observe that $\delta \cdot \mu = \sqrt{3n^\alpha \log n} \in o(n^\alpha)$. Therefore, our choice of δ provides the desired concentration $n^\alpha \pm \sqrt{3n^\alpha \log n}$. □

Lemma 5.2.2. *Let G be the social network described in Lemma 5.2.1. Divide the threshold interval $[0, n)$ into $M = k \cdot n^{1-\alpha}$ disjoint bins $B_i, i \in \mathcal{B}$, with threshold range n^α/k, where k is a constant positive integer. Each bin contains $n^\alpha/k \pm O(\sqrt{n^\alpha \log n})$ nodes.*

Proof. This proof is analogous to that of Lemma 5.2.1. Let Y_i, for $i \in \mathcal{B}$, be the random variable indicating the number of nodes in each bin i. With this change in bin size, the new combined union bound and Chernoff bound is:

$$\begin{aligned} \Pr\left[\bigvee_{1\le i\le M}\left(|Y_i - \mu/k| \ge \delta\mu/k\right)\right] &\le \sum_{1\le i\le M} \Pr\left[|Y_i - \mu/k| \ge \delta\mu/k\right] \\ &\le 2M \cdot e^{-\mu\delta^2/(3k)}. \end{aligned}$$

Next, we redefine the δ of Lemma 5.2.1, so $\delta = \sqrt{\frac{3k\log n}{n^\alpha}}$. Therefore,

$$\begin{aligned} 2M \cdot e^{-\mu\delta^2/3} &= 2 \cdot k \cdot n^{1-\alpha} \cdot \exp\left(-\frac{\mu\,\delta^2}{3}\right) \\ &= 2 \cdot k \cdot \exp\left((1-\alpha)\log n - \frac{n^\alpha\delta^2}{3}\right) \\ &= 2 \cdot k \cdot n^{-\alpha} \to 0. \end{aligned}$$

Finally,

$$\delta \cdot \mu = \sqrt{\frac{3n^\alpha \log n}{k}} \in O(\sqrt{n^\alpha \log n}).$$

□

5.3 Inducing Contagion

With our concentration results, we can finally resolve one of the principal questions of our research. Given a social network of n individuals, represented by a complete graph, what is a sufficient condition for the size of our seed to ensure contagion?

Theorem 5.3.1. *Let G be the social network described in Lemma 5.2.1 and let $2/3 < \alpha \leq 1$. If our seed consists of the n^α nodes with the highest thresholds, then contagion occurs* whp.

Proof. For some positive constant integer k,

$$M_k = k \cdot n^{1-\alpha}.$$

Partition $[0, n)$ into intervals of size n^α/k, so that the ith interval is

$$J_i = \left[(i-1) \cdot \frac{n^\alpha}{k}, i \cdot \frac{n^\alpha}{k}\right), \text{ where } 1 \leq i \leq M_k.$$

B_i denotes the ith bin, which is the set of nodes whose thresholds are in J_i. By Lemma 5.2.2, each bin is simultaneously concentrated *whp*.

When $i \cdot n^\alpha/k$ nodes are already active, we can *activate* bin i. This is because the number of active nodes will exceed the highest threshold represented in bin i. We will show that if the seed consists of the n^α nodes with the highest thresholds, then, *whp*, contagion will occur by inducing a domino effect on the set of bins $\mathcal{B}$.

The seed $\Phi(0)$ is contained in the bins with the $k+1$ highest indices, or the "last $k+1$" bins. Further, the overwhelming majority of those nodes will be in the last k bins. To understand the domino effect used in our proof, suppose that the first $K-1$ bins are currently active for some positive integer K. To activate all nodes in B_K, it is sufficient to have

$$\max_{j \in B_K} \theta_j \leq K \cdot \frac{n^\alpha}{k} \leq |\Phi(0)| + \sum_{i=1}^{K-1} |B_i|. \tag{5.1}$$

This inequality suffices as an activation condition, even in the worst case scenario of every node $v_j \in B_K$ possessing the highest possible threshold. With this criteria in mind, we describe our domino effect.

Case: $\alpha = 1$

This case is trivial as the seed consists of the entire network, $\Phi(0) = \mathcal{V}(K_n)$. Therefore, all nodes are activated at $t = 0$.

Case: $2/3 < \alpha < 1$

Consider the case $\alpha = 2/3 + \epsilon$ for some constant $0 < \epsilon < 1/3$. For this, we assume a division of the threshold interval according to $M_2 = 2n^{1-\alpha}$. Obviously, we can

activate bin B_1 because $\frac{n^\alpha}{2} < \Phi(0) = n^\alpha$. Next, we can activate bin B_2 *whp* because:

$$\max_{j \in B_2} \theta_j \prec n^\alpha < n^\alpha + \frac{n^\alpha}{2} - \sqrt{\frac{3n^\alpha \log n}{2}} \leq |\Phi(0)| + |B_1| \text{ whp.}$$

For $3 \leq K \leq M_2 - 1$, suppose the first $K-1$ bins are activated, along with our seed. This means that *whp* the total number of active nodes satisfies

$$\begin{aligned}
|\Phi(0)| + \sum_{i=1}^{K-1} |B_i| &\geq n^\alpha + (K-1)\left(\frac{n^\alpha}{2} - \sqrt{\frac{3n^\alpha \log n}{2}}\right) \\
&> \frac{K}{2} n^\alpha + \frac{1}{2} n^\alpha - K\sqrt{\frac{3n^\alpha \log n}{2}} \\
&> \frac{K}{2} n^\alpha + \frac{1}{2} n^\alpha - M_2\sqrt{\frac{3n^{2/3+\epsilon} \log n}{2}} \\
&= \frac{K}{2} n^\alpha + \frac{1}{2} n^{2/3+\epsilon} - 2n^{1/3-\epsilon}\sqrt{\frac{3n^{2/3+\epsilon} \log n}{2}} \\
&= \frac{K}{2} n^\alpha + \frac{1}{2} n^{2/3+\epsilon} - 2n^{2/3-\epsilon/2}\sqrt{\frac{3 \log n}{2}} \\
&> \frac{K}{2} n^\alpha \geq \max_{x \in B_K} \theta_x
\end{aligned}$$

for n sufficiently large.

Therefore *whp* we can activate bin B_K. Here, we note that to finish activating the entire network, there are three scenarios to consider.

In the first scenario, the total sum of nodes in the two highest bins is exactly n^α. Therefore, the last two bins contain no nodes other than those in the seed, and the seed is fully contained within the last two nodes. Our domino process will therefore stop at $K = M_2 - 2$. Once that bin is activated, we have activated the entire network because the last two bins were activated as part of the seed.

The second scenario is the case that our seed $\Phi(0)$ is fully contained in the last two bins, but the size of the seed is exceeded by the sum of nodes contained in these top two bins. Because of the concentration on our bins, it must be that the last bin only contains nodes in the seed. Therefore, there are some nodes in the second highest bin that are not members of the seed. Our domino process will continue to $K = M_2 - 1$, which it is valid for. While there will be fewer nodes activated as a result of this step than in previous steps, this is irrelevant as the last bin is activated as part of the seed and thus did not need the "push" towards activation from this second to last bin.

Finally, the size of our seed may slightly exceed the total number of nodes contained in the two highest bins. In this case, this slight "rollover" will be wholly contained in the third highest bin. This fact is directly obtained from the concentration of our bins. We then continue our process until $K = M_2 - 2$. As in the previous scenario, we will be able to activate this without an issue. However, we will activate less nodes than expected, because some were already

active as part of our seed. Since the following two bins are also already active as part of the seed, this is irrelevant as they did not need the "push" towards activation.

In each scenario, we have proven contagion *whp*.

□

This result tells us that a seed in $o(n)$ can be the **critical mass** which induces contagion on our network. We note that the result may still be dissatisfactory. Because our condition for contagion is sufficient but not necessary, a smaller lower bound may exist for the size of our seed that can be uncovered using different methods. At this point, it is important to note several features of this model and theorem.

- Our results are overly cautious, assuming several worst case scenarios: i) each node has the highest threshold possible within their bin, ii) each activated bin contained the least possible amount of nodes it could have *whp*, and iii) the deviation from the mean number of nodes was often overcompensated when multiplied by the maximum number of bins instead of just those activated. A more refined technique may yield contagion for some $\alpha < 2/3$.

- We focus on yielding contagion *whp*. In many applications, practitioners may be concerned merely with exceeding some given percentage chance. This would likely allow for reducing the parameter α, though such research would likely involve a different method of analysis.

- This model both assumes a complete network and equal weighting of all nodes. However, particularly within social networks, individuals weight the behavior of certain neighbors more strongly than others.

This final point is the main focus of our next chapter, as we explore how varying the weight of neighbors can affect our seed parameter α.

Chapter 6

Contagion in the Presence of Homophily

With analysis of diffusion in large, complete networks in place, we now consider dynamics in the presence of homophily. Homophily is defined as the tendency of individuals to associate with others similar to themselves. Within our model, we generalize this idea and define **homophily** as a node's disproportionate weighting (large or small) of nodes in certain groups in the network over others.

In the presence of homophily, we again redefine several characteristics of our original Linear Threshold Model.

6.1 A Weighted Linear Threshold Model

Definition 6.1.1 (Population). The **population** of our network, $\mathcal{V}$, is the union of k disjoint sub-populations:

$$\mathcal{V} = \bigcup_{i=1}^{k} V_i.$$

This definition directly implies that every individual belongs to some group within the social network. With this in mind, we can now introduce the idea of weighted influence. For any node in a group G_j, the influence on the node by some other node in the group G_i is denoted by $w_{i,j}$, whose value we currently leave unrestricted. Therefore, the total possible influence of neighbors on the node is equal to

$$\sum_{i \neq j} w_{i,j}|G_i| + w_{j,j}\left(|G_j| - 1\right)$$

and so the threshold vector in a graph K_n is a mapping

$$\theta_k : k \in G_j \rightarrow \left[0, \sum_{i \neq j} w_{i,j}|G_i| + w_{j,j}\left(|G_j| - 1\right)\right). \tag{6.1}$$

Note that the range of this threshold function depends on the subpopulation G_j and its influence coefficients $w_{i,j}$.

Because we have changed our threshold function, we must also present revised definitions for cohesive sets and the activation of nodes to reflect these new values.

Definition 6.1.2. A set of individuals $\mathcal{M} \subseteq \mathcal{V}$ form a cohesive set on the complete graph if, for any node v_i belonging to group G_j,

$$\sum_{k \neq j} w_{k,j}|G_k| + w_{j,j}|G_j| - \sum_{k} w_{k,j}|\mathcal{M} \cap G_k| < \theta_i \quad \text{for all} \quad v_i \in \mathcal{M}. \tag{6.2}$$

Definition 6.1.3. For a complete graph K_n, given a set $\Phi(t)$ of active nodes at time t, a node v_i in group G_j is active at time $t+1$ according to the following condition:

$$\sum_{k} w_{k,j}|\Phi(t) \cap G_k| \geq \theta_i \Rightarrow v_i \in \Phi(t+1) \tag{6.3}$$

While the additional notation complicates, the diffusion of a behavior is very similar to the process used thus far in our analysis. A weighted modification of Example 2.2.4 is presented below.

Example 6.1.4. Let our social network be represented by the complete graph K_5. We treat each individual as their own group V_i. Let the weighting of an individual be **consistent** among all players, such that $w_{i,j} = w_i$ for all $V_i \in \mathcal{V}(G)$. These weights are provided by the vector $w = [3, 2, 1, 4, 5]$. Each player's threshold is chosen uniformly at random, leading to the threshold vector $\theta = [1, 3, 9, 1, 2]$. From our values w and θ, we note that V_5 is the most influential group (and node) while V_3 is both the least influential and "stubborn." The following figure represents this network, where each player v_i is represented by their tuple (θ_i, w_i).

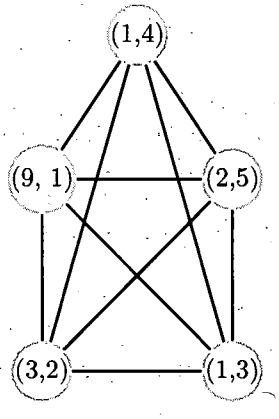


Figure 6.1: A social network represented by K_5.

If at the beginning of the process, player v_2 is designated as the seed, diffusion proceeds as follows:

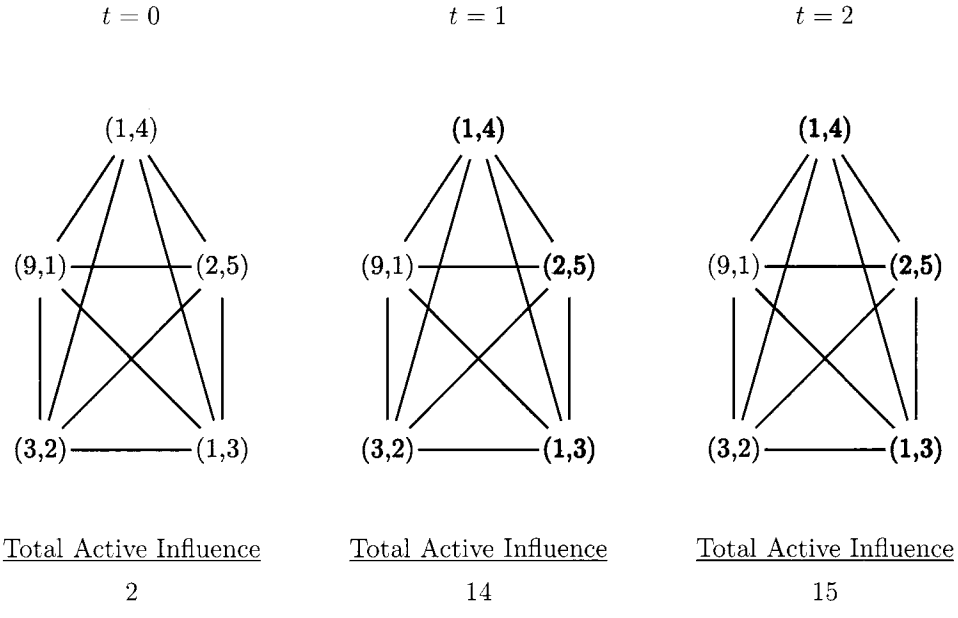


Figure 6.2: The diffusion process for a homophilous network on K_5.

Diffusion therefore converges and the modeled behavior has been adopted by the entire network by the end of $t = 2$.

One of the most important features of this weighted model is its clear mapping to the case of an unweighted model. In Example 6.1.4, we treated each node as its own group. Therefore, a mapping to the unweighted model could be provided by interpreting each node with the tuple (θ_i, w_i) as a group of w_i nodes, each with weight 1 and threshold θ_i. A seed can then be interpreted as activating the entire group formed by the previously single node. Reevaluating the diffusion process would yield the groups of nodes in this model being activated in the same order that nodes were activated in the weighted model. Our network from Example 6.1.4 is mapped in this way in Figure 6.3

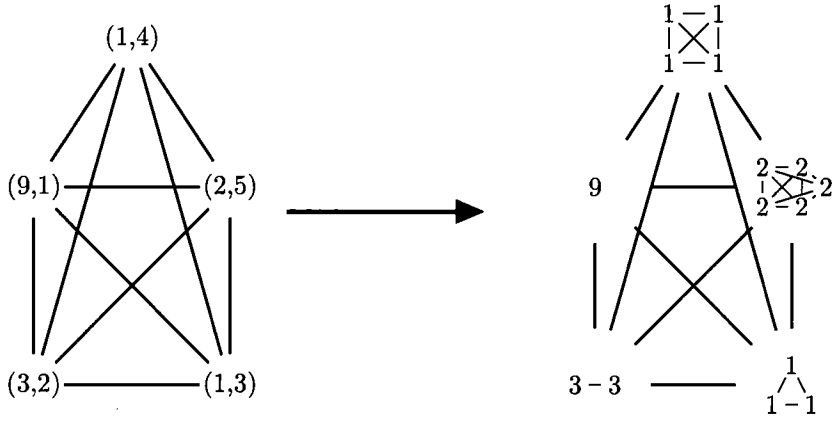


Figure 6.3: A mapping from the network K_5 to the unweighted network K_{15}.

We can generalize this to any case of the weighted model. When weighting is not consistent in a network, we provide a slight modification by mapping each group i to its perceived network N_i.

Definition 6.1.5 (Perception). Let V_i be a group of vertices in the social network represented by graph K_n. The group V_i has a **perception** N_i. This perception is a mapping from the original social network to how the network is perceived by members of the group V_i as an unweighted graph.

Within the perception, groups are mapped to nodes in a manner similar to Figure 6.3. Each group V_j is of size $w_{j,i} \cdot |V_j|$ and each node in these resized groups has unit weight. Therefore, N_i is represented by the graph with $\sum_j w_{j,i}|V_j|$ vertices. In this new graph, we note that each member of V_i keeps the same threshold. However, we do not care about the thresholds of any group other than V_i, nor how the nodes in these other groups are connected. It only matters that V_i form a clique, and that each node in each group V_j has an edge connecting it to each node in V_i. This allows us to have an unweighted model of the external influence on V_i.

If we are interested in seeing how many members of a different group, such as V_j, are activated by any time t during this process, we revert to the perception N_j to do so. We will note changes in group V_i while up until time $t-1$ in perception N_i. These changes are then reflected in the ratio of nodes from V_i activated in N_j.

Under this interpretation of the weighted Linear Threshold Model, a node in V_i is activated when the number of active nodes N_i. By providing this mapping, results in unweighted networks can be extended to homophilous networks.

With our new weighted model, we can now determine the effect of homophily on inducing contagion. If our social network is partitioned into a set of groups,

or *cliques*, does this increase or decrease the necessary size of our seed to induce contagion *whp*? We ask this question in two general cases. First, under what circumstances can a small group of highly influential individuals (such as politicians, celebrities, and other public figures) act as the seed for an idea to become contagious? We designate this small and highly influential group as **thought leaders**. Second, given a segmented network consisting of thought leaders and a general population, when can an idea grow from its seed, amongst the general population, and induce contagion?

To study these, we provide more formal definitions of the aforementioned social groups. We present two models of thought leaders, beginning with the case of leaders unaffected by the general population.

6.2 Immutable Thought Leaders

For a social network represented by a complete graph K_n, let T designate the group of thought leaders, where

$$|T| = f(n) \in o(n) \quad \text{and } f(n) \in \omega(1).$$

Let all other individuals in the network belong to a general population P such that

$$|P| = n - f(n)$$

and

$$\mathcal{V} = T \cup P.$$

Because of their status, each thought leader places a weight of 1 on other thought leaders. Because our groups here are named explicitly instead of begin indexed, we abuse notation and state that $w_{T,T} = 1$. Immutable thought leaders consider the general population as "just another voice" and thus having weight $w_{P,T} = |P|^{-1}$. This implies that the corresponding unweighted model for group T is a graph $K_{|T|+1}$ consisting of all thought leaders and a single individual representing the population. Theorem 6.2.1 below shows why we call these thought leaders "immutable."

The disparity of influence is similar amongst the general population. Thought leaders have a much stronger influence on a member of the general population than their peers in group P. However, the general population still consider themselves as important signals of a behavior's merits. Therefore, we set $w_{P,P} = 1$ and $w_{T,P} = \frac{n}{f(n)} = \frac{n}{|T|}$. These weights are presented as Table 6.1 below.

Influencers \ Influencees	T	P
T	$w_{T,T} = 1$	$w_{T,P} = \frac{n}{\lvert T\rvert}$
P	$w_{P,T} = \frac{1}{\lvert P\rvert}$	$w_{P,P} = 1$

Table 6.1: Weights in a model of immutable thought leaders

Under this model, a member of the general population considers P and T as equally important groups, and could be influenced by a seed from either set. This is because, as our network of size n grows, each group will have approximately equal influence on P. To understand this, we note that the total influence of group T on a node $v_i \in P$ is

$$\frac{w_{T,P} \cdot |T|}{w_{P,P} \cdot (|P| - 1) + w_{T,P} \cdot |T|} = \frac{\frac{n}{f(n)}(f(n))}{(n - f(n)) + \frac{n}{f(n)}(f(n))} = \frac{n}{2n - f(n)}.$$

Because $f(n) \in o(n)$, asymptotically

$$\frac{n}{2n - f(n)} \to \frac{1}{2}$$

meaning thought leaders have half of the influence on the general population. Further, there are only two groups so the proportion of total influence of the general population on an $v_i \in P$ is $1/2$ as well.

With these weights, we already understand the mapping of the way thought leaders and the general population perceive the world to unweighted models. We abuse notation to designate perceptions by group names instead of by index, as in N_T.

Theorem 6.2.1. *In a social network K_n comprised of a general population P and immutable thought leaders T, as $n \to \infty$,*

$$\Pr[\textit{ Contagion } |\Phi(0) \subseteq P] \to 0.$$

Proof. We conduct this proof by assuming the most generous of circumstances. Let $\Phi(0) = |P|$. We map the model to a perception N_T represented by $K_{|T|+1}$. In N_T, the group P is represented by a single node. Because this single node is connected to all members of T and the group T forms a clique, N_T is a complete graph.

Recall that the mapping in Lemma 4.1.4 which enumerated the cases where there were no non-trivial cohesive sets excluded the node with the lowest threshold. Therefore, on a graph K_{n+1}, it only mapped where the remaining n nodes placed their thresholds into n bins. Because θ_P does not exist in this network,

its null value is consistent across mappings of the remaining n nodes in N_T into the n bins their thresholds can exist in. Just as the node P cannot influence the formation of a cohesive set in N_T, the lowest threshold node could not influence the formation of a cohesive set because it always rested in the first bin.

Because of this consistency in not effecting the formation of cohesive sets, we can state that C_n enumerated the number of distribution classes without cohesive sets in N_T. Next, recall that this mapping also reflected the cases where any singleton seed could induce diffusion, including the excluded node with the lowest threshold. Because the node was excluded, whether or not it had a threshold did not affect our bijection; all that mattered was that if it was activated first, it could induce diffusion. Our node P in N_T can act as this excluded node in the mapping, where it has no threshold. It follows from Corollary 4.2.2 that the probability that this singleton seed P in N_T can induce diffusion is

$$\frac{1}{n} \to 0.$$

Finally, we note this acts as a bound on our probability. If the seed was a strict subset of the group P, the odds of success would be reduced because we would first need to determine the odds that this subset of P was large enough to activate all remaining members of P. Therefore,

$$\Pr[\text{ Contagion } |\Phi(0) \subseteq P] \leq \frac{1}{n} \to 0.$$

□

As expected, as $n \to \infty$, it becomes impossible to induce immutable thought leaders to adopt a behavior, no matter how large a portion of the general population has done so. Meanwhile, these thought leaders still hold considerable influence over the general population.

Theorem 6.2.2. *In a social network K_n comprised of a general population P and immutable thought leaders T, contagion occurs* whp *from $\Phi(0) \subseteq T$ when $\Phi(0) = (f(n))^\alpha$ for $2/3 < \alpha \leq 1$.*

Proof. We divide our proof into two cases, using the first as part of our justification in the second.

Case: $\alpha = 1$

If $\alpha = 1$ and $\Phi(0) \subseteq T$, then $\Phi(0) = T$. With all thought leaders active, we focus our attention on the general population, P. If all thought leaders are active, in the perception $N_P = K_{2n-f(n)-1}$ of group P, it is equivalent to

$$|T| \cdot \frac{n}{|T|} = n$$

nodes being active at time zero. We can then state that for N_P,

$$|\Phi(0)| = n \in \Theta(n).$$

Next, we note that the number of vertices in N_P is

$$m = 2n - f(n) - 1 \in \Theta(n).$$

Theorem 5.3.1 can immediately be generalized to state that for a network of size $\Theta(m)$, contagion occurs *whp* if $|\Phi(0)| \in \Theta(m^\alpha)$ for any $2/3 < \alpha \leq 1$.

Using this generalization, because $|\Phi(0)| \to \frac{m}{2} > m^{2/3}$, it meets the requirement of Theorem 5.3.1 in the special case of $\alpha = 1$. Contagion therefore occurs.

Case: $2/3 < \alpha < 1$

If $\alpha < 1$, then $\Phi(0) \subsetneq T$. Therefore, we first focus on contagion in the perception of thought leaders $N_T = K_{|T|+1}$. Using our generalized version of Theorem 5.3.1, we state that for a network of size $\Theta(|T|)$, contagion occurs *whp* if $|\Phi(0)| \in \Theta(|T|^\alpha)$. Because $w_{T,T} = 1$,

$$\Phi(0) = (f(n))^\alpha = |T|^\alpha \in \Theta(|T|^\alpha)$$

in perception $K_{|T|+1}$ as well. Therefore, within N_T, we meet the criteria for contagion and so every member of T is eventually active by some time k.

Next, we treat time k as $t = 0$ in the second step of our model. With all of T active, we focus on activating P. However, with $\Phi(0) = |T|$, we are presented with the scenario in the case of $\alpha = 1$. Therefore, we have already proven that each member of P will eventually be activated, and thus contagion occurs.

□

These results have demonstrated two important facts in societies with immutable thought leaders. First, it is impossible to change the behavior of thought leaders when they are not a member of the seed. This is because, in their perception, thought leaders clearly form a cohesive set. Since there is only a single node, P, the criteria for a cohesive set $\mathcal{M} = T$ is easily met:

$$\left(|T| + |P|^{-1}|P|\right) - |T| = 1 < \theta_i \text{ for all } v_i \in \mathcal{M}.$$

Second, thought leaders are an incredibly effective way to induce contagion in a network. While previously our seed needed to be of size $\omega(n^{2/3})$, we can now induce diffusion from $\omega(f(n)^{2/3})$ given a group of $f(n)$ thought leaders, no matter how small. However, in certain situations, these models can still be wanting, particularly if thought leaders in a network are not perceived to be immutable. In the following section, we loosen our characterization of thought leaders to account for such scenarios.

6.3 Flexible Thought Leaders

We again divide our total population of n nodes into a general population P and a group of thought leaders T, where $|T| = f(n) \in o(n)$ and$f(n) \in \omega(1)$. In this

section, we turn our attention to how much we can manipulate these weights. How little influence can thought leaders have on the general population while having $\Phi(0) \subseteq T$ still induce contagion? How much influence can the general population P have on the thought leaders in order for a seed $\Phi(0) \subseteq T$ to induce contagion in perception N_T? Is this influence sufficient for a seed $\Phi(0) \subseteq P$ to induce contagion on the thought leaders?

We begin with the first question. We first adapt Theorem 5.3.1 to our needs.

Lemma 6.3.1. *Let a network represented by K_n. Each group V_i in the network has size $|V_i|$ where*

$$|V_i| \in O(n),$$

$$|V_i| \in \omega(1),$$

and

$$\sum_i |V_i| = n.$$

If under the perception of N_i the seed $\Phi(0)$ meets the criteria

$$|\Phi(0)| \in \omega\left(\sum_j w_{j,i}|V_j|\right)$$

then all of V_i is activated.

Proof. The total influence on a node in group V_i is equal to the number of nodes in its perception N_i by construction. Therefore, the total number of nodes in N_i is

$$\sum_j w_{j,i}|V_j|.$$

To activate all of V_i, it is sufficient to induce contagion in the network formed by N_i. By Theorem 5.3.1, for K_m, if

$$|\Phi(0)| \in \omega\left(m^{2/3}\right)$$

then contagion will be induced *whp*. Because N_i is a complete graph network of size

$$\sum_j w_{j,i}|V_j|,$$

a seed

$$|\Phi(0)| \in \omega\left(\sum_j w_{j,i}|V_j|\right)$$

will induce contagion in N_i, thereby activating each node in V_i as desired. □

Theorem 6.3.2. *Assume a network consisting of a general population P and thought leaders T such that $\mathcal{V} = T \cup P$. If the relative weight of thought leaders on the general population is*

$$\frac{w_{T,P}}{w_{P,P}} \in \omega\left(\frac{n^{2/3}}{|T|}\right),$$

and $\Phi(0) \subseteq T$, then contagion can induced under the criteria of Lemma 6.3.1.

Proof. We begin by assuming without loss of generality a unit weight on self-perception in group P, e.g. $w_{P,P} = 1$. Given

$$|P| = n - |T|,$$

for any node $v_i \in P$, the total influence on v_i from group P is

$$n - |T| - 1 \in \Theta(n).$$

For contagion under the criteria of Theorem 5.3.1, it must hold that in perception N_P,

$$|\Phi(0)| \in \omega(n^{2/3}).$$

Thus, our aim is for

$$w_{T,P} \cdot |T| \in \omega(n^{2/3})$$

so that, once all of T is active, it can act as the seed in perception N_P. It follows from this condition that, given any group of thought leaders such that $|T| = f(n)$, contagion under the criteria of Lemma 6.3.1 occurs if

$$w_{T,P} \in \omega\left(\frac{n^{2/3}}{T}\right).$$

Under this condition, if a seed can activate all of T, then T will have sufficient influence to act as the seed for N_P, completing our proof. □

Now that we know the needed influence of thought leaders, we determine how much we can stretch the effects of the general population.

Theorem 6.3.3. *Assume a network consisting of a general population P and thought leaders T such that $\mathcal{V} = T \cup P$. If the relative weight of the general population on thought leaders is*

$$\frac{w_{P,T}}{w_{T,T}} \in O\left(\frac{f(n)}{|P|}\right)$$

and $\Phi(0) \subseteq T$, contagion can be induced under the criteria of Lemma 6.3.1.

Proof. We wish to assure that contagion occurs in perception N_T so that we can state that, from our seed $\Phi(0) \subseteq T$, each node in T is eventually active. Once T is active, we assume a weight in accordance to Lemma 6.3.2 and then can activate all nodes.

We again assume, without loss of generality, a unit weight on self-perception such that $w_{T,T} = 1$. Let

$$|T| = f(n) \in o(n)$$

where $n = |\mathcal{V}|$. For any node $v_i \in T$, the total influence on v_i from group T is

$$f(n) - 1 \in \Theta\left(f(n)\right).$$

In order for a seed $\Phi(0) \subseteq T$ and

$$|\Phi(0)| \in \omega\left(\left(f(n)\right)^{2/3}\right)$$

to meet the criteria of Lemma 6.3.1 for activating all of T, the total possible influence on $v_i \in T$ must be $\Theta\left(f(n)\right)$. The total influence on $v_i \in T$ is defined as

$$w_{T,T} \cdot (|T| - 1) + w_{P,T} \cdot |P| = f(n) - 1 + w_{P,T} \cdot |P|.$$

We know

$$f(n) - 1 \in \Theta\left(f(n)\right).$$

Thus, we only need to ensure that

$$w_{P,T} \cdot |P| \in O\left(f(n)\right)$$

because

$$\Theta\left(f(n)\right) + O\left(f(n)\right) \in \Theta\left(f(n)\right).$$

Therefore, our weighting requirement, given $w_{T,T} = 1$, is expressed as

$$w_{P,T} \in O\left(\frac{f(n)}{|P|}\right).$$

□

If the general population has a weighted influence exceeding the requirement in Theorem 6.3.3, then the total influence on group T will exceed $\Theta\left(f(n)\right)$, Therefore, a seed

$$|\Phi(0)| \in \omega\left(\left(f(n)\right)^{2/3}\right)$$

may not be sufficiently large in N_T for activating all of T, as it will not meet the criteria of Lemma 6.3.1.

Corollary 6.3.4. *Assume a network consisting of a general population P and thought leaders T such that $\mathcal{V} = T \cup P$. If the relative weight of the general population on thought leaders is*

$$\frac{w_{P,T}}{w_{T,T}} \in \omega\left(\frac{\left(f(n)\right)^{2/3}}{|P|}\right)$$

and $\Phi(0) \subseteq P$, contagion can be induced under Lemma 6.3.1.

Proof. We wish to ensure that if contagion occurs in N_P, and thus all of P is activated, P can serve as a seed for contagion in N_T. Thereby, all of T will be activated and contagion will have occurred in our network. Once again, we assume unit weighting on self-perception without loss of generality.

We have already noted that the total influence in perception N_P is $\Theta(n)$ in a network of size n. Therefore, assuming $|\Phi(0)| \in \omega(n^{2/3})$, the contagion criteria for N_P under Lemma 6.3.1 is satisfied. All of P is active. Because $|P| \in \Theta(n)$, this means there exists a seed $\Phi(0) \subseteq P$ that can activate all of P, completing the first step in our requirement.

With all of P active, we must now activate all of T. First, we restart time at the present point, such that $\Phi(0) = P$. In the network from Theorem 6.3.3,

$$w_{P,T} \in O\left(\frac{f(n)}{|P|}\right)$$

and so the total influence of the seed is

$$|P| \cdot O\left(\frac{f(n)}{|P|}\right) = O\left(f(n)\right).$$

However, in this network, contagion is induced if the total influence of the seed is $\omega\left((f(n))^{2/3}\right)$. Thus, we also need for

$$|P| \cdot w_{P,T} \in \omega\left((f(n))^{2/3}\right) \Rightarrow w_{P,T} \in \omega\left(\frac{(f(n))^{2/3}}{|P|}\right).$$

□

These results imply that there exists a model where thought leaders can still place very little value on the general population P but still adopt a behavior originating from a seed consisting solely inside P. How small is this influence? If we again assume a unit weight on self-influence, we have proven that

$$w_{P,T} \in \omega\left(\frac{(f(n))^{2/3}}{|P|}\right) = \omega\left(\frac{f(n)^{2/3}}{n}\right)$$

permits contagion. Because $f(n) \in o(n)$ and $|P| \in \Theta(n)$, this means a weighting $w_{P,T} \in o(1)$ can still permit contagion, which is an astonishing result on how a population weighed as having such little importance can still change behavior.

6.4 Alternative Clique Structures

While these results have been illustrative, we are not always looking to model networks consisting of only two groups, one of which are thought leaders. Instead, we may be dealing with networks possessing a general population and some k other groups. Each of these groups may be flexible, but insular, as in

our last section. Further, each may have enough of an influence to act as a seed for other groups, or the general population. Our previous results are directly applicable to these cases. In any case of k groups for any positive integer k, our asymptotic analysis forgoes concerns regarding coefficients and behaves identically. Through two such cases, we show how the model presented in Section 6.3 can be extended in this manner.

Example 6.4.1. Let our social network be defined by a general population P and two groups, T_1 and T_2, of sizes $f(n), g(n) \in o(n)$, respectively. We assume that each group forms a clique, or a complete subgraph. To understand how groups are connected, we state that two groups G_1 and G_2 in a network are **connected** if the subgraph consisting of G_1 and G_2 forms a complete graph. By this definition, in our network, P is connected to T_1 and T_2, but T_1 and T_2 are not connected. With each group modeled as a node, Figure 6.4 displays the graph of our network of groups.

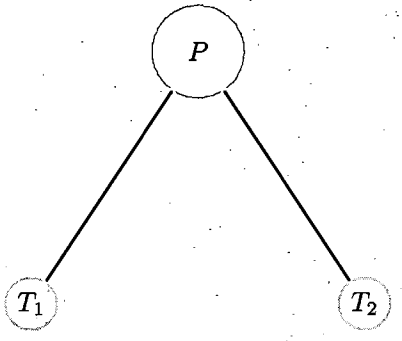


Figure 6.4: A social network with two disconnected cliques

Further, assume that all the relative weight requirements from Section 6.3 are preserved between each clique and the general population. Because Corollary 6.3.4 is preserved, as P is the only group influencing either T_1 or T_2, we do not need to modify any of our restrictions to state that a seed $\Phi(0) \subseteq P$ can induce contagion throughout the network.

However, could a seed existing in either T_1 or T_2 induce contagion? This case is of slightly greater complexity, as group P's perception includes influence both from the clique containing the seed and from the other. Assume once again that each group has a unit weight of self-influence. Further, assume without loss of generality that $\Phi(0) \subseteq T_1$. By preservation of the weight in Theorem 6.3.2,

$$w_{T_1,P} \in \omega\left(\frac{n^{2/3}}{|T_1|}\right).$$

Because the total influence on a node $v_i \in P$ from group P remains $\Theta(n)$ and the added influence of each clique remains sublinear. Therefore, the total possible influence on a node $v_i \in P$ remains:

$$n - |T_1| - |T_2| \in \Theta(n) - o(n) - o(n) = \Theta(n).$$

We know from preserving the population's weight in Theorem 6.3.3 and the isolation of T_1 from T_2 that a seed

$$|\Phi(0)| \in \omega\left((f(n))^{2/3}\right)$$

can activate all of T_1. Once all of T_1 is active, we reset time to $t = 0$. The influence of T_1 on P is then

$$|T_1| \cdot \omega\left(\frac{n^{2/3}}{|T_1|}\right) = \omega\left(n^{2/3}\right).$$

Thus, our seed is large enough to activate all of P. With all of P active, we have already shown that we can instigate all of a clique such as T_2 to be active.

This example serves as an excellent case of the following Theorem.

Theorem 6.4.2. *Let a social network of n nodes be defined by $k+1$ complete subgraphs: a general population P and k cliques T_i of size $f(n) \in o(n)$. P is connected to each clique T_i, but all T_i are disconnected. If $k \in \mathbf{N}$ and the relative weight restrictions from Theorem 6.3.2, Theorem 6.3.3, and Corollary 6.3.4 are preserved, such that*

$$\frac{w_{T,P}}{w_{P,P}} \in \omega\left(\frac{n^{2/3}}{|T|}\right),$$

$$\frac{w_{P,T}}{w_{T,T}} \in O\left(\frac{f(n)}{|P|}\right),$$

and

$$\frac{w_{P,T}}{w_{T,T}} \in \omega\left(\frac{(f(n))^{2/3}}{|P|}\right),$$

respectively, contagion can be induced from a seed formed by a subset of any clique T_i or P and large enough to first activate its entire clique.

Proof. The proof for this Theorem is analogous to Example 6.4.1 and is a direct consequence of our use of asymptotic analysis. □

However, we may be more interested in social networks in which groups form a complete graph, as opposed to those with a tree structures. Will these results hold when all cliques are connected as well? For such a generalization, we must rethink what we mean by preservation of the relative weight restrictions from Theorem 6.3.2, Theorem 6.3.3, and Corollary 6.3.4. We note two patterns amongst our weight restrictions:

- the weight of any group G_i on G_j must be large enough to act as a seed if all of G_i is activated;
- the weight of any group G_i on G_j must be small enough so that the order of total influence on G_j is unchanged.

With these considerations in mind, we redefine our weight restrictions of any group G_i on G_j to the following:

$$\frac{w_{G_i,G_j}}{w_{G_j,G_j}} \in \omega\left(\frac{(|G_j|)^{2/3}}{|G_i|}\right) \tag{6.4}$$

$$\frac{w_{G_i,G_j}}{w_{G_j,G_j}} \in O\left(\frac{|G_j|}{|G_i|}\right) \tag{6.5}$$

and use them for the following corollary.

Corollary 6.4.3. *Let a social network K_n be defined by $k+1$ complete subgraphs known as groups. If $k \in \mathbf{N}$ and the relative weight restrictions from* (6.4) *and* (6.5) *are preserved, such that*

$$\frac{w_{G_i,G_j}}{w_{G_j,G_j}} \in \omega\left(\frac{(|G_j|)^{2/3}}{|G_i|}\right)$$

and

$$\frac{w_{G_i,G_j}}{w_{G_j,G_j}} \in O\left(\frac{|G_j|}{|G_i|}\right),$$

contagion can be induced from a seed formed by a subset of any group when it is large enough to first activate its entire group.

Proof. The proof for this corollary is also direct consequence of our use of asymptotic analysis and the generalization to (6.4) and (6.5). For any k, the total possible influence on a group G_i remains at the same order because of (6.5) insisting each new group G_j have a sublinear influence. Because the original order was the order of $|G_i|$, a seed of size $\omega(|G_i|^{2/3})$ may still active all of G_i.

Once all of any group G_j is activated, it can activate any other group by (6.4). □

In this section, we have studied how we can extend the homophily model to two different types of social networks: networks with groups forming graphs of single generation trees, and groups forming complete graphs. We believe that these two models are widely applicable. Groups forming trees can act as basic models of understanding the flow of ideas in nations with several rural populations disconnected from one another. If an idea captures the attention of one such group (such as a new agricultural technology) news may spread to a general population which then disseminates it to other rural communities.

Similarly, complete graphs of groups represent more flattened social structures. A society may have no general population, and instead be comprised of k different demographics which all influence each other but hold their own group in greater regard. Once contagion occurs in the isolation of any one group, it may take off in all other groups. In these networks, our results appear to be qualitatively consistent with the study of diffusion in SIS models within Jackson & López-Pintado (2012).

Despite the applicability of these models, there are various other network structures and definitions of homophily that may be of interest, particularly in applied settings. We consider these in Chapter 7.

Chapter 7

Future Research

In this paper, we found a lower bound for the size of a seed needed to be the critical mass for contagion within the Linear Threshold Model. We began with small networks, exploring the odds of contagion for a seed of size k, for any constant positive integer k. Using the probabilistic method, we found that in large networks of size n, the lower bound on our seed is $|\Phi(0)| \in \omega\left(n^{2/3}\right)$. This was reduced when we introduced homophily, showing that an influential group of size $f(n) \in o(n)$, when fully activated, can act as a successful seed for diffusion in the entire network, our lower bound can be reduced to $|\Phi(0)| \in \omega\left(f(n)^{2/3}\right)$.

Our research, however, focuses on a narrow class of networks and diffusion models. There are several open problems that remain for interested readers to explore. The first of these lies in an implementation of random graphs. Determining the probability of contagion from small seeds on small random graphs, extending the work of Watts (2000) and ourselves, would provide valuable first steps in obtaining realistic probabilities for marketing and development practitioners interested in very small communities. More advanced random graph models in large networks would also be useful for providing insights on how connectivity i) affects the lower bound on our seed and ii) develops a boundary condition for determining if nodes for our seed should be determined by threshold value θ_i or measures of node centrality.

There are also opportunities for new models of homophily. In addition to introducing different graph structures or weighting structures, random graph models incorporating homophily would allow redefinition of homophily based on the connectivity of different groups. Reexamining our results in these models could be useful, and allow more direct comparison to empirical research. One potential model for homophily is the Block Two-Level Erdős-Rényi model (Seshadhri et al., 2012), which provides close approximations to a number of real-world networks.

Finally, a number of other diffusion models exist. Some models study diffusion in the presence of competing behaviors, or activate an individual based on probabilistic interactions with a single other individual at any given time. A study of lower bounds on seeds for contagion could prove useful in each of these

circumstances. One such model that we propose is a generalization of the Linear Threshold Model where thresholds θ_i are drawn independently at random from an arbitrary probability distribution in (0, 1]. This line of research is inspired by work determining the distribution of thresholds in social networks, including Valente (1996) who finds a variety of distributions including both those that are near uniform and near bimodal. Bimodal distributions could be of particular interest in random graphs when determining the relative importance of picking seeds with high thresholds and those with high centrality.

Bibliography

Acemoglu, D., Ozdaglar, A., & Yildiz, E. (2011). Diffusion of innovation in social networks. In *Proceedings of the 50th IEEE Conference on Decision and Control* (p. 2329-2334). Institute of Electrical and Electronics Engineers.

Adler, J. (1991). Bootstrap percolation. *Physica A, 171*.

Berger, E. (2001). Dynamic monopolies of constant size. *Journal of Combinatorial Theory Series B, 83*, 191-200.

Blume, L. E. (1993). The statistical mechanics of strategic interaction. *Games and Economic Behavior, 5*.

Blume, L. E. (1995). The statistical mechanics of best-response strategy revision. *Games and Economic Behavior, 11*.

Brock, W. A., & Durlauf, S. N. (2000). *Interactions-based models* (Working Paper No. 258). National Bureau of Economic Research.

Domingos, P., & Richardson, M. (2001). Mining the network value of customers. In *Proceedings of the Seventh International Conference on Knowledge Discovery and Data Mining.* Association for Computing Machinery.

Domingos, P., & Richardson, M. (2002). Mining knowledge-sharing sites for viral marketing. In *Proceedings of the Eighth International Conference on Knowledge Discovery and Data Mining.* Association for Computing Machinery.

Dupas, P., & Cohen, J. (2010). Free distribution or cost-sharing? Evidence from a randomized malaria prevention experiment. *Quarterly Journal of Economics, 125*, 1-45.

Gleeson, J. P., Melnik, S., Ward, J. A., Porter, M. A., & Mucha, P. J. (2012). Accuracy of mean-field theory for dynamics on real-world networks. *Physical Review E, 85*.

Granovetter, M. (1978). Threshold models of collective behavior. *The American Journal of Sociology, 83*.

Jackson, M. (2010). *Social and Economic Networks.* Princeton University Press.

Jackson, M., & Golub, B. (2009). *How homophily affects learning and diffusion in networks* (Working Paper No. 35.2009). Fondazione Eni Enrico Mattei.

Jackson, M., & Golub, B. (2012a). Does homophily predict consensus times? Testing a model of network structure via a dynamic process. *Review of Network Economics*, *11*.

Jackson, M., & Golub, B. (2012b). How homophily affects the speed of learning and best-response dynamics. *The Quarterly Journal of Economics*, *127*.

Jackson, M., & López-Pintado, D. (2012). *Diffusion and contagion in networks with heterogeneous agents and homophily* (Working Paper No. 2012/12). Université catholique de Louvain, Center for Operations Research and Econometrics.

Janson, S., Luczak, T., Turova, T., & Vallier, T. (2012). Bootstrap percolation on the random graph $G_{N,P}$. *The Annals of Applied Probability*, *22*.

Kempe, D., Kleinberg, J., & Éva Tardos. (2003). Maximizing the spread of influence through a social network. In *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Association for Computing Machinery.

Kempe, D., Kleinberg, J., & Éva Tardos. (2005). Influential nodes in a diffusion model for social networks. In *Proceedings of the Thirty Second International Colloquium on Automata, Languages, and Programming*. Springer Lecture Notes in Computer Science.

Kleinberg, J. (2007).
In N. Nisan, T. Roughgarden, Éva Tardos, & V. V. Vazirani (Eds.), *Algorithmic Game theory* (chap. Cascading Behavior in Networks: Algorithmic and Economic Issues). Cambridge University Press.

Koshy, T. (2009). Lobb's generalization of Catalan's parenthesization problem. *The College Mathematics Journal*, *40*.

Lobb, A. (1999). Deriving the nth Catalan number. *Mathematical Gazette*, *83*.

McPherson, M., Smith-Lovin, L., & Cook, J. M. (2001). Birds of a feather: Homophily in social networks. *Annual Review of Sociology*, *27*.

Mitzenmacher, M., & Upfal, E. (2005). *Probability and Computing: Randomized Algorithms and Probabilistic Analysis*. Cambridge University Press.

Morris, S. (2000). Contagion. *Review of Economic Studies*, *67*.

Rogers, E. M. (1962). *Diffusion of Innovations*. Free Press.

Schelling, T. (1978). *Micromotives and Macrobehavior*. Norton.

Seshadhri, C., Kolda, T. G., & Pinar, A. (2012). Community structure and scale-free collections of Erdős-Rényi graphs. *Physical Review E, 85*.

Stanley, R. (2001). *Enumerative Combinatorics* (Vol. 2). Cambridge University Press.

Stanley, R. (n.d.). *Solutions to exercises on Catalan and related numbers.* `http://www-math.mit.edu/~rstan/ec/catsol.pdf`. (Accessed: 2013-07-01)

Valente, T. W. (1996). Social network thresholds in the diffusion of innovations. *Social Networks, 18*.

Watts, D. J. (2000). *A simple model of fads and cascading failures* (Working Paper No. 2000-12-062). Santa Fe Institute.

Yildiz, E., Acemoglu, D., Ozdaglar, A., & Scaglione, A. (2011). Diffusion of innovation on deterministic topologies. In *Proceedings of the 36th International Conference on Acoustics, Speech and Signal Processing.* Institute of Electrical and Electronics Engineers.