COMPLEX CONTAGION IN SOCIAL NETWORKS

A Dissertation

Presented to the Faculty of the Graduate School

of Cornell University

in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

by

Christopher J. Cameron

May 2016

COMPLEX CONTAGION IN SOCIAL NETWORKS

Christopher J. Cameron, Ph.D.

Cornell University 2016

This work improves the applicability of complex contagion theory to empirical data by examining the assumptions of previous models of complex contagion and methods for measuring theoretically relevant quantities from observational data.

I consider the dynamics of complex contagion processes on networks with different distributions of tie range. Following prior work I explore the impact of permuting networks to induce more long range ties but also examine the contagion motifs and larger scale dynamics observed during the contagion life-cycle. I identify a mechanism—the formation of persistent locally isolated clusters—that triggers sudden mid-life-cycle increases in the rate of contagion on networks with some long range ties. I argue that both the number and the distribution of long range ties is critical to development of such clusters.

A survey of tie range distributions in empirical social networks revealed a consistent qualitative pattern where the range of long range ties is decidedly shorter than previous complex contagion models assumed.

I examine contagion dynamics on networks with more realistic distributions of tie range. Even though the neighborhood to node contagion motifs are substantially the same on both types of networks, realistic tie range distributions virtually eliminate the endogenous development of the clusters. Contagion on these networks is more robust to permutation but does not produce the mid-cycle accelerations found in networks with uniformly random permutation.

Finally, I identify a measurement issue that biases estimates of node thresholds from observed exposure at activation. I show the bias can be reduced by collecting additional data about non-activating exposures and using only a subset of observations to produce the estimates.

The empirically informed models presented in this work provide new insight about the potential outcome and dynamics of complex contagion processes in human social networks. Improved estimation techniques for node thresholds allow particular cases of empirical contagion to be described in terms of complex contagion theory. The results support an ongoing research program in complex contagion.

BIOGRAPHICAL SKETCH

Christopher J. Cameron received a BA degree in Sociology from the California State Polytechnic University, Pomona and a Ph.D. in Sociology from Cornell University. As a graduate student, he was an IGERT Fellow in the Cornell University IGERT Program in Nonlinear Systems and assistant director of the Social Dynamics Laboratory.

ACKNOWLEDGEMENTS

I would like the members of my committee: Michael Macy, Douglas Heckathorn and Jon Kleinberg for their guidance and inspiration. I would also like to acknowledge my office-mates Vladimir Barash and George Berry who shared their insights. I have benefited from the talented and diverse members of Michael Macy's Social Dynamics Laboratory — I can imagine no better colleagues. Finally, I am thankful for the support and patience of my family while I pursued my Ph.D. work.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

INTRODUCTION

1.1 Overview

The present work builds on my research effort for Barash, Cameron, and Macy (2012) and develops complex contagion theory to better bear on empirical data. The results presented here support a collaborative and ongoing research program in complex contagion with the Social Dynamics Laboratory at Cornell University.

In Chapter 1, I describe the literature relevant to the complex contagion model with a focus on recent results.

In Chapter 2, I consider the dynamics of the complex contagion model in the presence of long range ties. I examine the contagion motifs and larger scale dynamics observed during the contagion life-cycle and explore the impact of permuting networks to induce more long range ties. I identify the mechanism that allows complex contagions to spread faster on networks with some long range ties, namely, the formation of persistent locally isolated clusters. I argue that both the number and the distribution of long range ties is critical to the development of such clusters.

In Chapter 3, I survey tie range in empirical social networks and identify a consistent qualitative pattern where the range of long range ties is decidedly shorter than previous models assumed. I propose an alternative permutation strategy that produces more realistic distributions of tie range.

In Chapter 4, I recast results from complex theory in terms relevant to empirical networks and examine contagion dynamics when the distribution of long range ties is more consistent with empirical observations. Even though the neighborhood to node contagion motifs are substantially the same on both types of networks, realistic tie range distributions virtually eliminate the endogenous development of the clusters. Contagion on these networks is more robust to permutation.

In Chapter 5, I identify a measurement issue that biases estimates of node thresholds from observed exposure at activation. I show the bias can be reduced by collecting additional data about non-activating exposures.

Finally, Chapter 6 discusses some limitations of the present work and suggests directions for future research.

1.2 Origins and applications of the complex contagion model

Social network diffusion is a topic of long standing interest to social science. The earliest well known example comes from Coleman, Katz, and Menzel (1966) where doctors who were more well connected to peers tended to be among the first to prescribe a new drug. As a research program that offers the potential to identify influential individuals and to predict the outcomes of diffusion processes, social network diffusion has attracted researchers from economics, communications, sociology, epidemiology and marketing, resulting in diversity of models and approaches.

Valente (1995) identifies several classes of diffusion models including rela-

tional and structural models. In Valente's framework, the complex contagion model considered here is a relational model in that the peer network is the primary source of information. More specifically, it is a threshold model which developed from the combination of Granovetter's threshold model of collective action and his work in the role of weak ties in information diffusion (Granovetter, 1978; Grannovetter, 1973; Granovetter, 1983).

1.2.1 Threshold Model

Granovetter (1978) argued that the distribution, interaction and aggregation of preferences in a population may explain collective action outcomes and proposed the threshold model of collective action. In particular, the model illustrates that two slightly different distributions in preferences can produce widely divergent outcomes in the otherwise identical processes.

In Granovetter's model, there is a population of actors who each have an activation threshold. The threshold is defined as the proportion of the population that must adopt before ego adopts. Activation thresholds might range from 0 to the size of the population — they may extend outside this range, but can be bounded to this interval with no change in outcome. The model works by forward propagation. At first, there are no activated actors. In the first time step, any actors with a threshold of zero activate. In each subsequent time step, any actor whose threshold is less than or equal to the number of currently activated nodes is activated. This process continues until the number of nodes active reaches an equilibrium value which may be between zero and the size of the population.

Recast in a network framework, Granovetter's 1978 model is based on a fully connected network of actors. While he discusses the impact of social structure, noting that each actor might not be influenced by all the other actors in the population, he models the social structure by sampling threshold from the population and creating a smaller, fully connected network to represent the set of influential actors. He considers the effect of tie strength by making a small subset of the edges — friendships — twice as influential and does not find any marked difference in outcome for the distributions tested.

The major insight from Granovetter's threshold model is that small differences in the distribution of thresholds can lead to different outcomes. Because people observe behavior and not preferences the collectively constructed outcome does not necessarily reflect the collective preferences of the actors. Granovetter suggests it may be model the diffusion of innovations, rumors, diseases, strikes, educational attainment, migration and other social outcomes. That the distribution of thresholds and the information available to reactive actors is a viable explanation for different collective outcomes represented a major theoretical innovation that continues to inspire new work.

1.2.2 Role of long ties in information diffusion

For an actor embedded in a network, their network neighborhood is a major source of information about the state of other actors. Discarding the assumption of a fully connected network, the network structure becomes an important part of the information aggregation mechanism in the threshold model. Though he did not consider the implications to the threshold model, Grannovetter (1973)

explored the impact of local bridges and tie strength on the spread of information. Granovetter defined sociological tie strength as a function of time, emotional intensity, intimacy and reciprocity inherent in the tie. He described local bridges as ties connecting two endpoints where the next shortest path between the end points was greater than two. Because the local bridge connect two nodes that do not have any other neighbors in common, any information passed between the endpoints is likely to be novel to the received node. Using sociometric data, Granovetter demonstrated that open triads are statistically rare and concluded that strong ties are therefore less likely to be local bridging ties than weak ties. Because information flows along ties in a social network, bridge ties are critical connections to remote sources of information. Granovetter (1983) supported this conclusion using survey data that showed weak ties are a source of information about jobs for white-collar workers. Though it is based on a sample of ego-networks and not a full network, this study offers compelling evidence that graph structure can influence the spread of information.

While the strength or weakness of a tie is determined by the relationship the tie represents, the range of a tie is a structural graph property. Tie range is measured by the length of the second shortest path between the endpoints. A *local* or *short range* tie is an edge between endpoints with a second longest path of length two. A *long range* tie corresponds to a local bridge — a tie between nodes who are connected by a second longest path of length greater than two (Grannovetter, 1973). The relational model of influence is embedded in a larger social network whose structure impacts the outcome of diffusion processes.

1.3 Social network structure

1.3.1 The paradoxical structure of social networks

Contemporary social networks tend to possess two structural properties that present something of a puzzle; the average path length between any two nodes is short relative to the number of nodes and edges in the graph but a large proportion of edges are short range ties. Holland and Leinhardt (1978) proposed a measure to detect statistically significant departures from random tie formation and evaluated a number of empirical social networks. They concluded that most social networks show significant non-random structure and most have much higher transitivity than would be expected by chance. Transitivity is directly related to short range ties as each tie in a closed triad is a short range tie. Later examinations of larger networks found that the clustering coefficient for social networks was generally orders of magnitude larger than would be expected if ties formed at random (Watts & Strogatz, 1998; Newman, 2003). Despite the preponderance of short range, locally clustered ties, the nodes in social networks are generally connected by short paths.

Milgram (1967) conducted a famous letter passing experiment that demonstrated the possibility of traversing the American social acquaintance network to deliver a letter from midwestern cities to a particular person in Cambridge, Massachusetts. In this study and a followup Travers and Milgram (1969), people were able to effective direct letters though the social network connecting starting nodes to endpoints using an average of five intermediate message passers. In other words, people were connected by a path of no more than six steps on average. Travers and Milgram (1969) note that there is no way to know if the

letters could have been passed more efficiently so the experiment provided an upper bound on path length. With the advent of large scale network data collection it is possible to measure the average shortest distance between nodes directly and several studies summarized by Newman (2003) have reported consistent results for mean geodesics with lengths ranging from 4 to 8. Even the largest online social networks with millions of users still have mean geodesics with lengths ranging from 4 to 6 (Mislove, Marcon, & Gummadi, 2007).

1.3.2 The small world network model

Watts and Strogatz (1998) introduced a perturbed ring lattice network model that reconciled the high clustering and short characteristic path typical of empirical social networks. Starting with a network that consisted of entirely local ties, they introduced new ties between random nodes. These perturbations were much more likely to create long range ties so the process resulted in local bridges that connected otherwise distant parts of the network. Most importantly, they demonstrated that only a small number of long range ties are needed to dramatically reduce the mean geodesic length, so the change to the local neighborhood of most nodes is very small and the network clustering coefficient remains virtually unchanged.

This small world model provides an intuition for the nature of social networks that accommodates both the preponderance of short range ties and the short average path length that can also be framed in terms of Granovetter's strong and weak ties (Grannovetter, 1973; Granovetter, 1983). As a demonstration of the impact of the permutation on average path length, Watts and Strogatz

(1998) showed the rate of spread of an infectious disease dramatically increases with small amounts of random rewiring. Interpreted in sociological terms, the presence of only a few weak ties is sufficient to dramatically increase the speed of information diffusion in networks.

1.4 Complex contagion on small world networks

Centola and Macy (2007) recognized that weak tie explanations implicitly overgeneralize social contagions and argued that some contagions would be better modeled by introducing adoption thresholds, particularly contagions involving strategic complementarity, credibility, legitimacy or emotional contagion. They introduced the concept of complex contagion, defined as a contagion that requires "contact with two or more sources of activation" for transmission to occur (707). Centola and Macy argued that the insights drawn from Grannovetter (1973) and Watts and Strogatz (1998) will not apply to complex contagion because the bridging ties that help spread simple contagions are, by definition, not transitive. Due to the lack of redundant exposure to any contagion transmitted via bridging ties, nodes are unlikely to receive the additional contacts required for transmission. Based on analysis of a ring lattice and simulation results for a randomly rewired grid lattice, Centola and Macy (2007) establish that complex contagion cascades on these networks are qualitatively different from simple contagion cascades. In contrast to simple contagions, there is no increase in the speed of contagion for small numbers of bridging ties, there is a modest speed increase with increasing large numbers of bridging ties up to a point and then a sudden increase in time to saturate followed by cascade failures after a critical upper limit of rewiring ties. This study established that complex conta-

gion cascades have distinct characteristics and that the spread of complex contagions critically depends on the existence of sufficiently wide bridges between neighborhoods of nodes.

Since the introduction of the complex contagion model Centola and Macy (2007), three studies have found empirical evidence for the model. Leskovec, Adamic, and Huberman (2007) reported an increasing probability of a consumer buying books, music or DVDs with each additional personal recommendation received from other users. Using an experimental design, Centola (2010) demonstrated that a user's probability of adopting of a health behavior increased with the the number of neighbors who previously adopted and that compared to the random networks used as a control, diffusion of this adoption behavior was facilitated by networks with more local clustering. Recent research analyzing the cascade dynamics of Twitter hashtags (Romero, Meeder, & Kleinberg, 2011), revealed that categories of hashtags differ in the probability of adoption at different exposure levels. In particular, the use of politically themed tags is associated with a higher exposure at adoption than other—potentially less risky—tags.

Weng, Menczer, and Ahn (2013) examined the propagation of hashtags on Twitter through clusters of nodes identified by the InfoMap and link clustering community detection algorithms. Compared to less frequently used tags, the most used hashtags tend to appear in more clusters during the initial spread of the contagion. In contrast, the initial uses of less successful tags tended to be localized within a single cluster of well connected users. They argue that viral tags tend to spread in a pattern more consistent with simple contagion and non-viral tags tend to spread in a pattern more consistent with complex contagion. These

results are consistent with the finding from Centola and Macy (2007) that simple contagions spread many times faster than complex contagions, so any fixed observation window will see more successful simple contagions than complex contagions.

Barash et al. (2012) examined the process of complex contagion in small world networks, identifying a second mode of complex contagion once a critical mass of nodes became infected. During the initial period of infection, complex contagion requires dense local ties in order to spread. As the initial cluster of infection grows, most newly susceptible nodes are connected to the infected nodes via short range ties. In the initial phase, infected nodes may expose neighbors with whom they share a long range tie, but this single source of exposure is not sufficient to trigger the infection. Suppose the contagion threshold is T, defined as the number of infected neighbors necessary to trigger a node to adopt. As the size of the infected cluster grows, the probability that some uninfected nodes will have T long range ties to the infected cluster increases. Barash et al. (2012) showed analytically that these nodes exist with high probability once a critical mass adopts and that the expected proportion of nodes susceptible entirely via long range ties increases dramatically with the number of infected nodes. Beyond this critical point, the complex contagion no longer requires wide bridges between neighborhoods. Barash et al. (2012) also identified a second type of probabilistic event that establishes a new cluster via long range ties that then grows by leveraging wide bridges to nearby neighborhoods. The possibility of cascade failure due entirely to the location and structure of the seed cluster is particularly notable because it may serve as a model for the boom or bust nature of memes, fads and certain products.

Several papers suggest that the spread of a complex contagion may proceed through phases. Centola, Eguíluz, and Macy (2007) identify a pattern of slow spread followed by a pattern of accelerating spread. Barash et al. (2012) describe local and non-local spread, associating the change in the rate of spread with an increase in the probability of contagion via long range ties. Ghasemiesfeh, Ebrahimi, and Gao (2013) describe a local and accelerating phase of contagion and Taylor et al. (2015) describe wave front propagation followed by the appearance of new clusters. Monitoring aspects of the contagion process might enable detection of a shift in the dynamics (Barash et al., 2012) or allow prediction about eventual outcomes Weng et al. (2013).

CHAPTER 2

COMPLEX CONTAGION DYNAMICS

2.1 The complex contagion model

In a simple model of behavioral contagion on a network, an individual ego adopts a behavior if at least T of their alters adopt the behavior. For a given ego and behavior, T is ego's activation threshold. The theory of complex contagion posits the existence of behavioral contagions for which many nodes have an activation threshold greater than one (Centola & Macy, 2007). For a complex contagion, where the node thresholds T are greater than one for a preponderance of nodes in the network, the pattern of contagion in the network is qualitatively different from a simple contagion process, where most node thresholds are less than one. As I will show, complex contagions can produce especially sudden bursts of adoption and these bursts are associated with the emergence of *locally isolated* clusters of adopters.

2.1.1 Previous findings

Prior work with the complex contagion model suggests that the speed at which a contagion saturates a network is influenced by structural properties of the network and that the dynamics of the process may depend on the number or proportion of infected nodes. Centola et al. (2007) noted different growth rates between complex contagions on unperturbed and perturbed lattice networks; the growth rate for a contagion is initially faster on an unperturbed lattice compared to a contagion with the same threshold on a small world lattice. After the

size of the infected cluster is sufficiently large, the rate of infection on the perturbed lattice increases rapidly until almost all nodes are infected. In contrast, the rate of infection decreases gradually on the unperturbed lattice. Despite the slower start, complex contagions on the perturbed lattice saturate the network in fewer time steps than contagions on the unperturbed lattice.

Centola and Macy (2007) identified local bridges between graph neighborhoods as a fundamental structural prerequisite for complex contagion on a lattice. They demonstrate analytically that the bridge width in a lattice determines the maximum threshold contagion that can saturate the network. When a lattice has sufficiently wide bridges, some fraction of edges are redundant for a given contagion threshold and the network is robust to permutation up to some critical level of rewiring. Randomization is far more likely to disrupt existing bridges than create new wide bridges, so beyond the critical level of rewiring, bridges between neighborhoods are no longer sufficient to support contagion. Though Centola and Macy (2007) primarily focused on the transition from a network structure that supports contagion to one that does not, also found that higher threshold contagions spread more slowly than lower threshold contagions. Furthermore, they found that moderate levels of randomization decrease the time to saturation for lower threshold complex contagions and either did not decrease or increased the time to saturation for thresholds closer to the maximum supported by the network. They noted the possibility that random ties can create bridges to multiple neighborhoods and that the potential for activation via these bridges would increase with the number of nodes activated.

Once permutation creates long range ties in a network, there is a possibility that some nodes will have a sufficient number of long range ties to activate en-

tirely via long range ties. The probability that such a node exists is a function of the contagion threshold, the proportion of long range ties and the current number of active nodes in the network. Barash et al. (2012) examined the probability of these events analytically finding that such nodes are extremely likely to exist in small-world networks once a critical mass of nodes adopt. This probability also increases rapidly from near zero before the critical point to near unity after the critical point. This jump in probability occurs over a very small change in the total number of infected nodes. For instance, the addition of 90 nodes or 0.2% of the network size is sufficient to move the probability from near zero to unity for a threshold two contagion (Barash et al., 2012, p. 456). The size of the critical mass depends on the contagion threshold T and the proportion of long range ties. For simple contagions, ($T = 1$) the size of the critical mass is always one node. For minimally complex contagions ($T \in \{2, 3, 4, 5\}$) on perturbed lattices with small world properties, the size of the critical mass is much smaller (0.025%–0.10%) than the number of nodes in the graph (Barash et al., 2012, p. 456). As the contagion threshold increases or the fraction of rewired ties decreases, the size of the critical mass quickly approaches the number of nodes in the graph.

While Barash et al. (2012) considered the possibility of activation entirely via long range ties, they note two other modes of complex contagion. Some nodes will activate due to a combination of long and short range ties, which makes their analysis conservative (Barash et al., 2012). Another possibility is that some nodes with sufficient long range ties will have neighbors in common. Once the long range ties activate these focal nodes, they can infect their neighbors via short range ties and establish a new infected neighborhood whose members do not have any short range ties to the original cluster of adopters.

Taylor et al. (2015) considers spatial networks, particularly the permuted ring lattice and the London transit network. They describe two phenomena observed in a complex contagion process; wave front propagation and the appearance of spatially remote clusters but the focus of the work was recovering information about the network structure by examining the spread of many contagions over a single network. The patterns they describe correspond to some of the elements of the contagion dynamics discussed later in this chapter.

In sum, previous research suggests complex contagion proceeds through qualitatively different phases driven largely by changes in the probability of certain events as the number of infected nodes increases. Next, I will use simulation results to explore the dynamics of complex contagion and highlight the importance of the emergence of *locally isolated* clusters of adopters in driving changes in the adoption rate.

2.2 Contagion dynamics

2.2.1 Simulations of complex contagion

I use simulation to demonstrate the properties of complex contagion. These simulation are based on a static network structure. A connected set of nodes serve as a seed cluster of active nodes —the mechanisms that generate seed clusters are beyond the scope of the model. Nodes are updated in random order. Each time an inactive node updates, it checks the activation status of its network neighbors and becomes active if T or more neighbors are active, where T is the activation threshold for the target node. As the simulation continues,

the contagion saturates the network or reaches a state where no new nodes can be activated. Simulation time is measured in the number of node updates required to saturate the network. Following Centola and Macy (2007) and building on the work in Barash et al. (2012), I use a perturbed regular lattice mapped to a torus. A regular lattice offers an environment with high triadic closure so complex contagion is possible. Like the ring lattice used in Watts and Strogatz (1998), a regular lattice can the perturbed to induce the small world properties of low diameter and high local clustering. The perturbation technique, sometimes called the double edge swap, was introduced by Maslov and Sneppen (2002). This approach has the benefit of preserving the degree of each node, thus preserving both the degree distribution and the density of the graph. A perturbed regular lattice could also be a simple model of a spatial network, where most edges are between nodes who are physically near each other. Using a spatial analogy, edges in the unperturbed lattice are local ties and any perturbed edges tend to connect nodes that are not physically proximate. Unless a significant fraction of the edges are perturbed, nodes are very likely to share a short range tie with their spatial neighbors. Any ties between nodes that are not physically proximate are almost certainly long range ties. The spatial analogy is particularly useful for visualizing contagion dynamics.

2.2.2 Contagion motifs and phases

In the initial phase of complex contagion, new adopters tend to be well connected to nodes in the original seed cluster. In the spatial representation, new adopters are physically proximate to prior adopters, forming the dense cluster labeled A in Figure 2.1. In this stage, contagion moves from neighborhood to

neighborhood via wide bridges as demonstrated by (Centola & Macy, 2007). As this original cluster increases in size, some number of nodes will have T ties to nodes in the original cluster via the long range perturbed ties, leading to locally isolated adopters, such as those labeled B in Figue 2.1. These nodes reflect the contagion motif examined in (Barash et al., 2012) and the prevalence of these nodes is influenced by the proportion of long range ties in the graph. Finally, a third pattern of complex contagion emerges when these locally isolated adopters have neighbors with a sufficient number of long range ties. These groups of locally isolated adopters form clusters (C in Figure 2.1). Though contagion via long range ties establishes new clusters, clusters grow mostly via local, short-range ties.

These new clusters originate from one of two configurations. In one case, T nodes with T long range ties to the infected cluster who have T neighbors in common. In the second case, a cluster forms around a single focal node with T long range ties and $T-1$ neighbors who have $T-1$ ties to the infected cluster. With sufficient connectivity among the neighbors of the focal node, each activated neighbor can contribute to the activation of the next neighbor. Consider the collection of long range tie counts to the infected nodes for the neighbors of the focal node. If $\{x : 1 \leq x < T, x \in \mathbb{Z}\}$ is a subset of the collection, then the neighbors of the focal node will create a new cluster of sufficient size to activate nearby neighborhoods via local bridges. For instance, to create a new cluster when $T = 3$, the focal node must have three ties to the infected cluster and it must have one neighbor with two ties and one neighbor with one tie to the infected cluster and these neighbors must have ties between them.

The emergence of clusters defines the next phase in spread of a complex con-

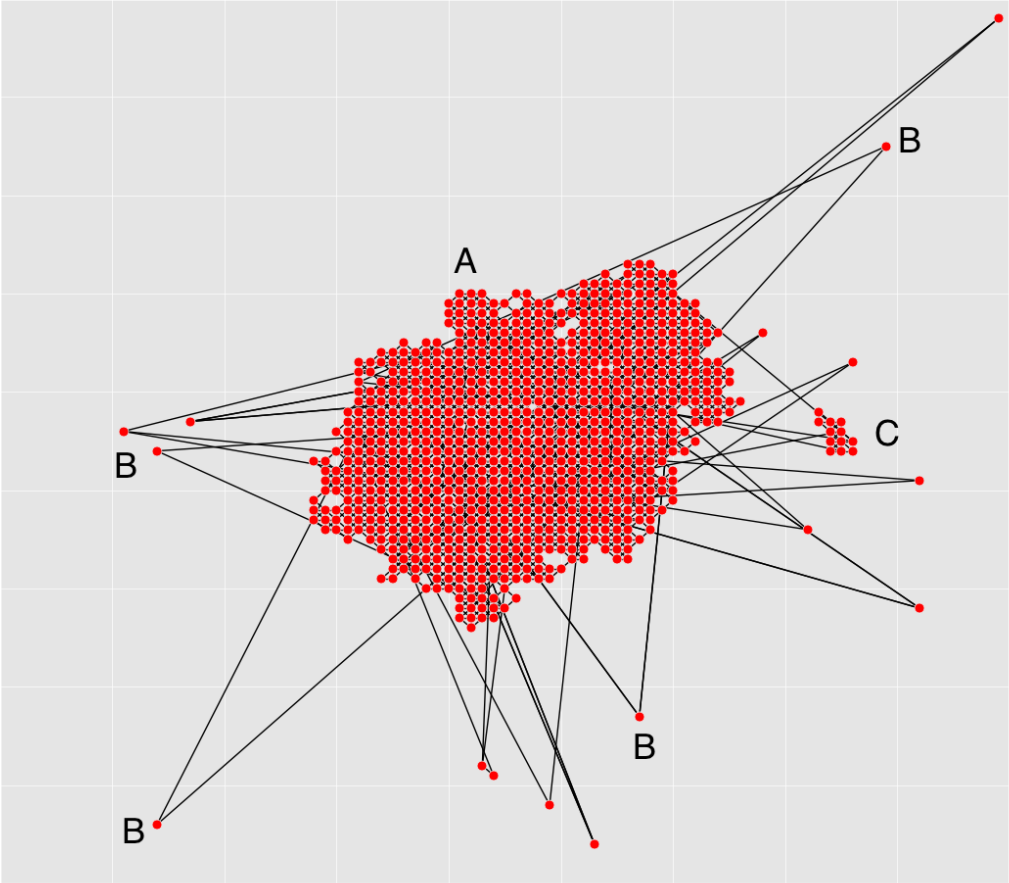


Figure 2.1: Elements of complex contagion: The large, well connected cluster (A); Several locally isolated nodes (B); and the beginning of a new locally isolated cluster (C). In this snapshot of a complex contagion process on a lattice, only the active nodes and the ties between active nodes are shown.

tagion. The clusters are only connected to the original seed cluster by long range ties, so it is a *locally isolated* source of exposure. A cluster exposes new parts of the graph, increasing the number of susceptible nodes. For a given number of infected nodes, multiple scattered clusters are much more efficient at exposing new nodes than one monolithic cluster. At the same time, there are many scattered, locally-isolated active nodes. While these nodes cannot provide the reinforcement needed to activate their network neighbors, they do effectively reduce T by one for each of their neighbors. Widespread exposure coupled

with reduced effective thresholds leads to dynamics more typical of simple contagions in the later parts of the contagion process.

The complex contagion occurs through three contagion motifs: neighborhood to neighborhood via wide bridges, neighborhood to neighborhood spread assisted by long range ties and the activation of locally isolated nodes via long range ties. An activation entirely via long range ties supported by a few long range assisted activations can establish a viable locally isolated cluster. These motifs are present to different degrees with different thresholds of contagions and different proportions of long range ties. In the next section, I will consider how these elements combine to produce life-cycle differences in contagions of different thresholds.

2.2.3 Life-cycle differences in simple and complex contagion

For simple contagions on networks with long range ties, any activation via a long range tie is sufficient to create a viable locally isolated cluster of new adopters. Since a locally isolated adopter is a viable locally isolated cluster for a simple contagion and only a single long range tie is necessary to establish each remote cluster, viable clusters are likely to appear early in the life-cycle. As described in Watts and Strogatz (1998), the long range ties extend the area of contagion to graph regions remote from the seed cluster and the contagion quickly exposes a large number of nodes. In contrast, locally isolated clusters are more difficult to establish with complex contagions. For complex contagions, the chance of establishing locally isolated adopters - and the chance that these adopters will form clusters - depends on the contagion threshold, the pro-

portion of long range ties and the number of current adopters. Differences in the probabilities of these events lead to striking qualitative differences in the longer term dynamics and life-cycle of complex contagions.

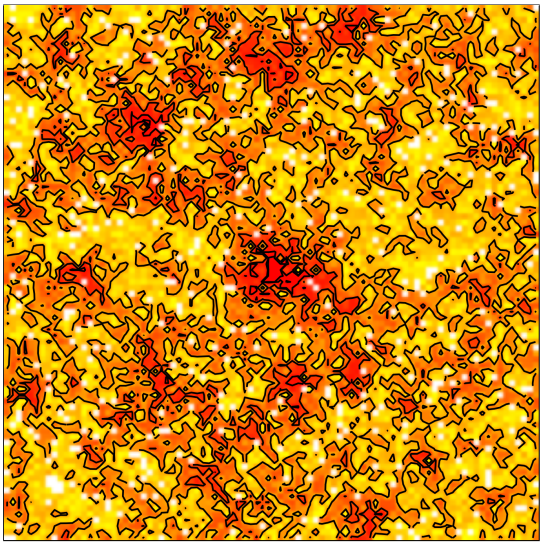
Pattern of contagion with different thresholds

Depending on the contagion threshold, locally isolated clusters may emerge at different points in the life-cycle or may fail to form entirely. Figure 2.2 shows the dynamics of contagion on a permuted 8-regular lattice of 8100 nodes with 10% long range ties. In each of the sub-figures, the network structure and seed cluster are exactly the same - the only difference is the contagion threshold. For ease of comparison, the seed cluster is centered in each image. The contour lines show the extent of adopters at 20, 100, 1000 and 4000 nodes adopted.

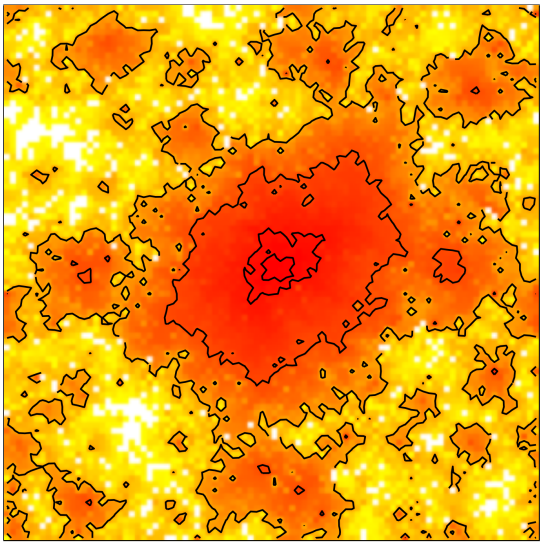
For a simple contagion, (Figure 2.2a) the order of adoption only loosely corresponds to the location of the initial seed cluster. The number of nodes exposed by prior adopters is large compared to the number of adopters.

For the complex contagions (Figure 2.2b,c) new adopters are localized around the seed cluster because contagion initially proceeds through wide local bridges. For the lower threshold complex contagion (b) locally isolated clusters eventually emerge and then serve as new sources for localized spread via local bridges. As these clusters emerge, the number of nodes susceptible relative to the number of adopters increases - clusters increase the efficiency of exposing new nodes. In contrast, thresholds near the capacity of the network (c) only rarely establish viable clusters of adopters and do so late in the life-cycle after many nodes have already adopted. Constrained to spread from neighborhood

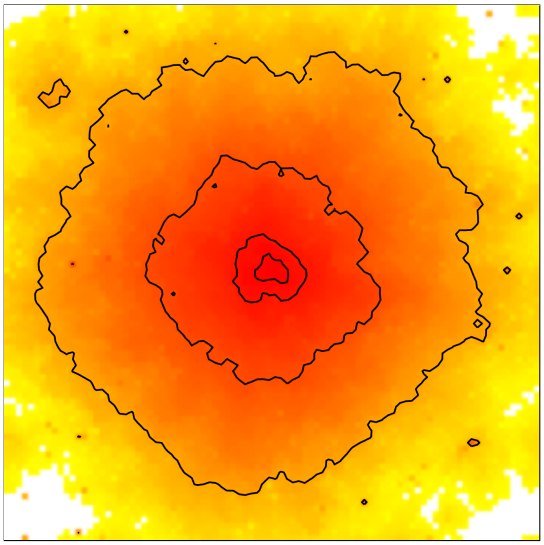
to neighborhood via wide bridges, these contagions radiate from the seed cluster.



(a) Threshold = 1



(b) Threshold = 2



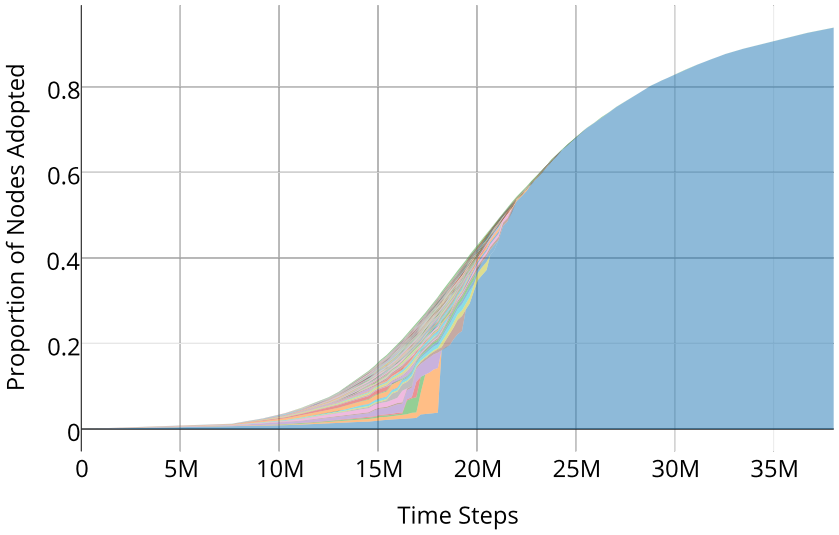
(c) Threshold = 3

Figure 2.2: The history of contagion on a spatial representation of a permuted lattice graph. Pixels represent individual nodes at corresponding coordinates lattice coordinates and pixel color reflects order of activation. Red nodes activated first with lighter colors activating last.

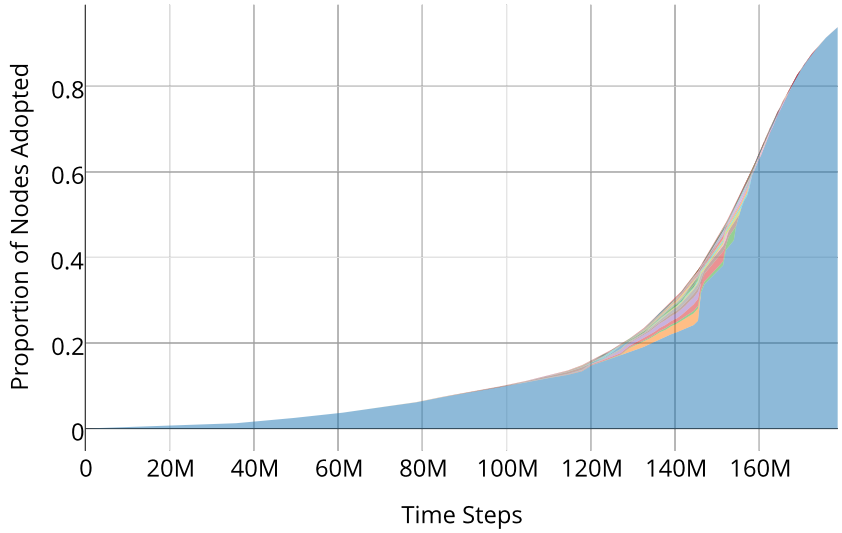
Cluster emergence and growth over the life-cycle

The preceding adoption time heat maps (Figure 2.2) rely heavily on the spatial representation of lattice graphs. Though useful to develop an intuition about the differences in the dynamics, the spatial representation is not useful for larger graphs nor even sensible for non-planar, non-spatial graphs. In order to extend the insight and explore other network structures it is necessary to track the emergence and growth of clusters over time without depending on visual inspection.

Tie range provides a graph-structure-based measure of distance between nodes that does not rely on the position of nodes in space. The range of an edge e is the length of the second shortest path between the endpoints of e. Locally isolated clusters are defined as groups of nodes connected to each other via short range ties. At a particular point in time during the contagion process, clusters can be identified by removing long range ties from the social graph and then inducing a subgraph on the set of nodes that have adopted the contagion. Each component in the subgraph is a locally isolated cluster. If clusters are labeled by the earliest adoption time among the nodes in the cluster, then clusters have a persistent identity between time points. When two clusters merge, the merged clusters have the same identity as the older cluster. Measuring the size of the clusters at each time point produces a timeline that shows when clusters emerge and how the population of adopters is divided among the clusters.



(a) Threshold = 1

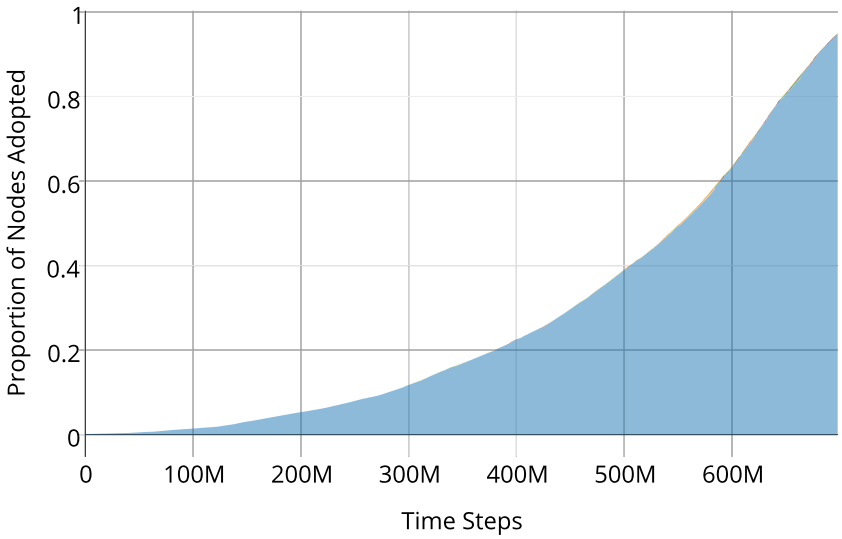


(b) Threshold = 2

Figure 2.3: Adoption of a contagion over time with clusters highlighted. The infected nodes are grouped into locally isolated clusters. The width of each colored band shows the fraction of nodes within that cluster at that time step. For complex contagions (b), the development of clusters is delayed and precedes a dramatic increase in the rate of adoption.

When broken down by locally isolated components, the adoption history (Figure 2.3) shows the same qualitative patterns apparent from visual inspection of the heat map (Figure 2.2) as well as some new features. In the adoption history for a simple contagion (2.3a) the clusters appear very early and the rate of adoption increases as new clusters emerge. As the contagion spreads, clusters begin to collide and the rate of adoption begins to decline until the network is saturated.

For minimally complex contagions (2.3b), the emergence of clusters is delayed. Before clusters emerge, the growth rate is nearly constant. Interestingly, the growth rate within clusters is also relatively constant which indicates growth patterns dominated by contagion via local bridges. The increase in overall rates of adoption can be explained as the additive effect of many new clusters emerging. Finally, the emergence of these clusters precedes a dramatic increase in the rate of adoption.



(a) Threshold = 3

Figure 2.4: Adoption of a threshold 3 contagion over time with clusters highlighted. The higher threshold contagion does not establish any sizable viable clusters.

Contagions with thresholds near the maximum supported by the contagion do not benefit from long range ties. Figure 2.4 shows that the rate of adoption increases gradually as the proportion of adopters increases. Clusters may appear late in the contagion life-cycle, but grow too slowly to dramatically increase the number of exposed nodes.

In the preceding sections, I have examined specific realizations contagion processes to illustrate features in the life-cycle of a complex contagion. The remained of the chapter examines how the life-cycle dynamics change as the proportion of long range ties varies.

2.3 Long range ties and complex contagion

2.3.1 Contagion motif dynamics

In Section 2.2, I described three contagion motifs that might be observed in a complex contagion, including the possibility of contagion entirely via long range ties. Barash et al. (2012) showed that beyond a critical mass of nodes, a complex contagion could spread entirely via long range ties between adopters and non-adopters. Even though contagion entirely via long range ties is possible, this motif does not necessarily dominate the later contagion life-cycle nor does it necessarily explain the sudden increase in the adoption rate. In this section I consider the relative frequency and dynamics of the different contagion motifs on networks with different proportions of long range ties.

To better relate the present work to the results from Centola and Macy (2007),

I replicated the findings relating the presence of long range ties to the number of iterations required for contagions of different thresholds to saturate a network. In contrast to Centola and Macy (2007) and Watts and Strogatz (1998), I parametrize the networks by the proportion of long range ties rather than the proportion of edges rewired to facilitate comparison with empirical social networks. Figure 2.5 shows that increasing the contagion threshold increases the number of iterations to saturation and that simple and minimally complex contagions benefit from the presence of some long range ties. Contagions near the maximum threshold supported by the network are impaired when some short range ties are rewired to create long range ties. Too much permutation is harmful for complex contagions. These results are consistent with those reported in Centola and Macy (2007).

The fraction of node adoptions attributable entirely to long range ties increases with the proportion of long range ties in the graph. It is worth noting in Figure 2.5 that the marginal benefit of additional long range ties is quite minimal in graphs where at least 10% of the ties in a graph are long range ties. Figure 2.6a shows the proportion of node activations attributable entirely to long range ties. Considering the window around 10% long range ties, it is clear that local spread —perhaps aided by some long range ties— is the dominant contagion motif. Observing a single contagion spread over a network in this window, we would see very few instances of contagion via long range ties.

Aside from the overall rarity of activation entirely via long range ties, there is little evidence that this motif occurs with more frequency later in the contagion life-cycle. Figure 2.6b shows the number of nodes active in the network at the median time of activation via a long range ties. If most of the activation

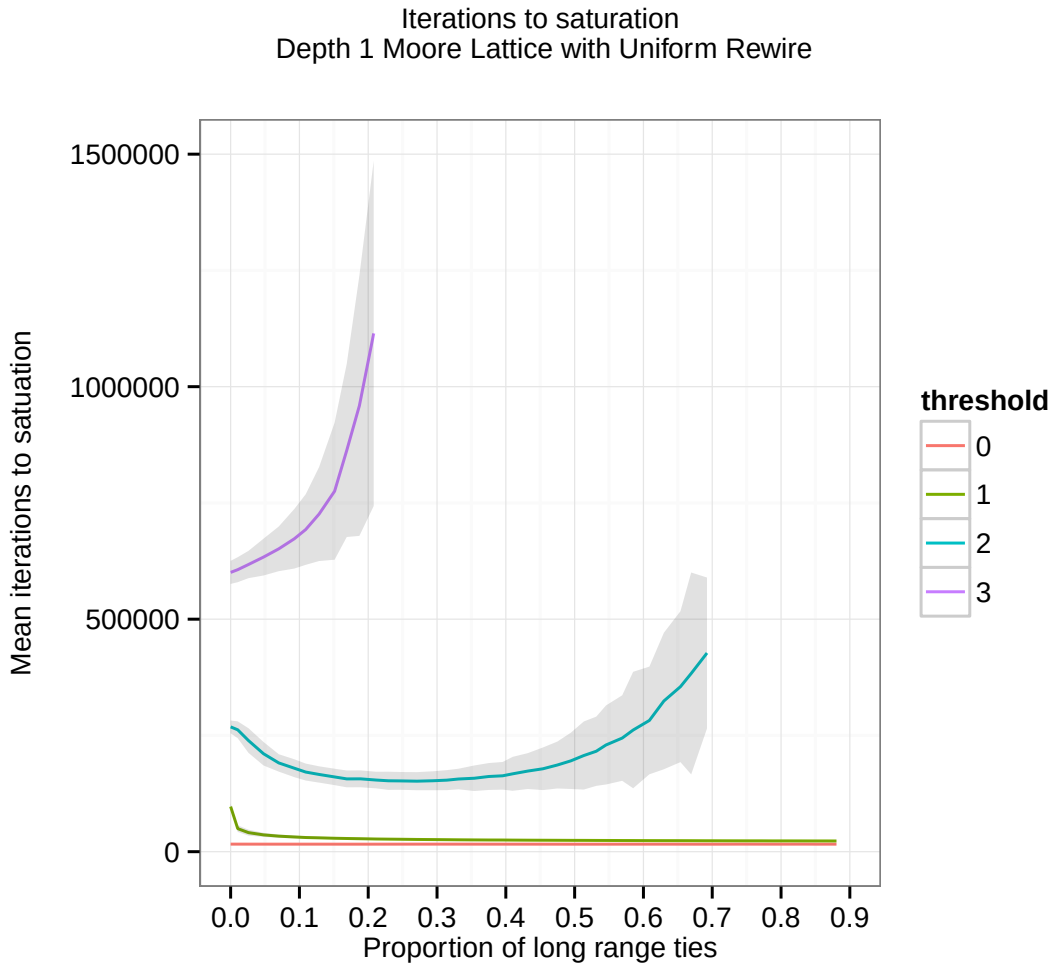
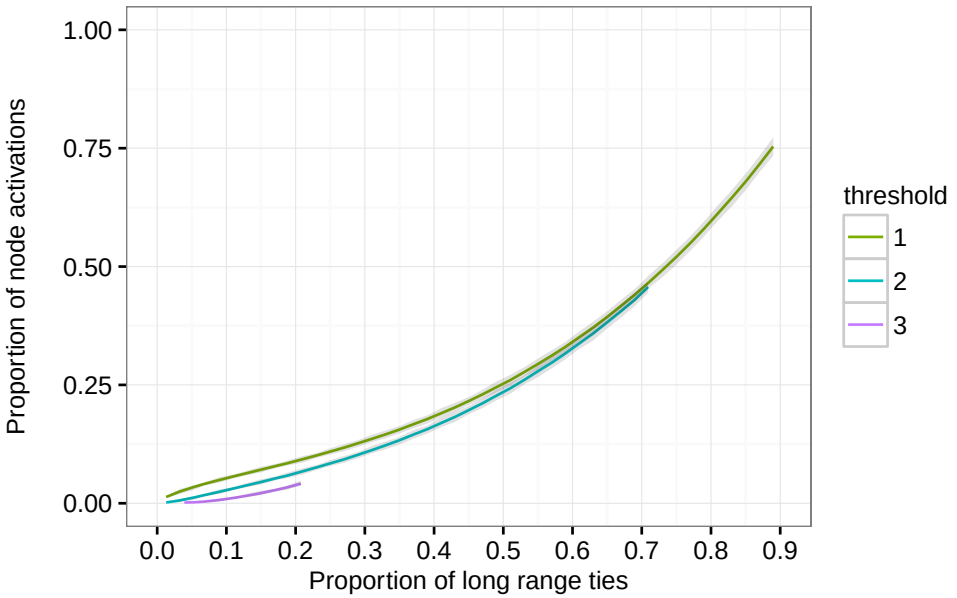


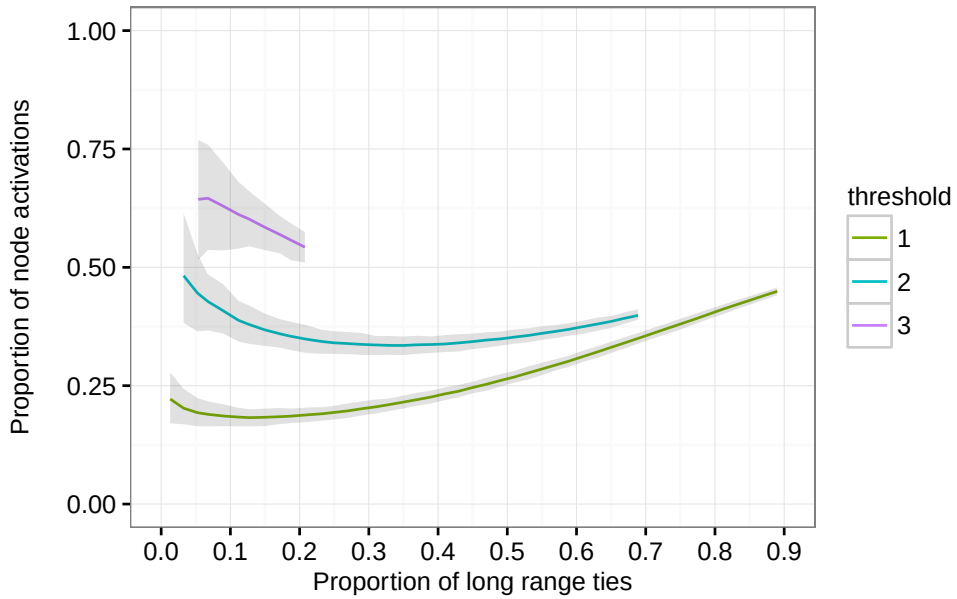
Figure 2.5: Mean iterations to saturation for contagions with various thresholds on a regular Moore lattice with degree 8.

via long range ties occurred towards the end of the final stages of contagion, then the median activation would occur at a higher proportion of nodes active. Instead, the median long range activation for simple and minimally complex contagions, $T = \{1, 2\}$, is before the median node activation. For contagions that benefit from long range ties ($T = 2$), activation via long range ties occurs throughout the contagion process. Ninety percent of the activations by long range ties occur between 10 and 75% of the nodes activated, with a somewhat higher event density in the first half of the contagion process.

Despite the relative rarity of activations via long range ties and the lack of a distinct change in the frequency of these events over the course of the contagion process, this contagion motif plays a critical role in decreasing the time required to saturate a network. Long range ties establish locally isolated adopters and a fraction these adopters serve as the focal node for new clusters. Clusters don't change the observed contagion motifs — local spread is the dominant motif throughout the contagion life-cycle.



(a) Proportion of nodes activated via long range ties.



(b) Proportion of nodes active at median long range activation time.

Figure 2.6: Timing and frequency of activation via long range ties for contagions of different thresholds on permutations of an 8-regular Moore lattice.

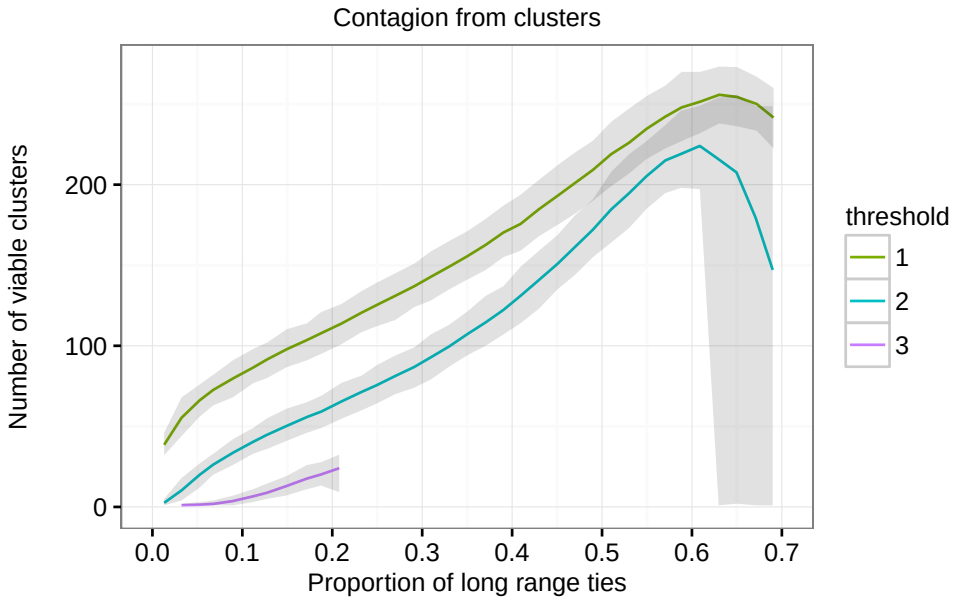
2.3.2 Cluster dynamics

The proportion of long range ties does have a marked impact on the formation of clusters and the change in cluster dynamics corresponds to the change in iterations to saturation observed by Centola and Macy (2007).

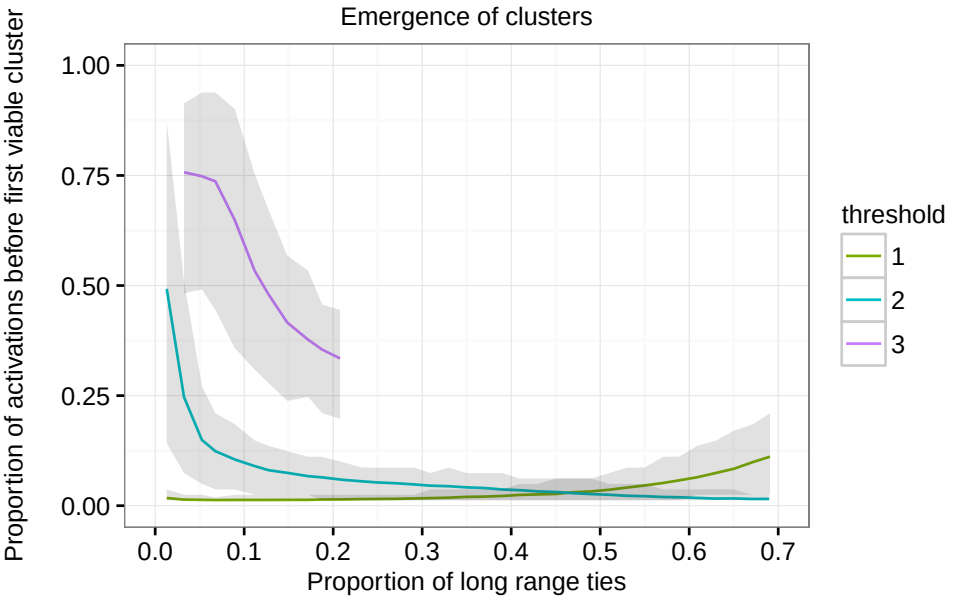
The number of viable clusters produced in the life-cycle increases linearly with the proportion of long range ties in the network. For the purposes of this analysis, a cluster with nine or more nodes is considered viable as the initial contagion starts from nine active nodes. In general, clusters of any size greater than the contagion threshold are viable so the number of clusters is fairly stable across cutoff values. In the interval between 0-10% long range ties, where additional long range ties, the number of seed clusters increases from 1 to dozens (Figure 2.7a). Since each cluster represents a new area of local spread, increasing the number of clusters increases the rate at which new nodes are exposed to the contagion.

The timing of non-seed cluster emergence is another important factor in saturation times. The results from Barash et al. (2012) virtually guarantee than long range activations will occur at some point in the contagion life-cycle if long range ties are present. The clusters that these activations spawn reduce saturation time more if they occur earlier in the life-cycle — a viable cluster at the 90% saturation cannot substantially decrease the overall time to saturation whereas a second cluster in the early contagion life-cycle could nearly double the contagion efficiency. Figure 2.7b shows that clusters emerge much sooner in the life-cycle as the proportion of long range ties increases, with the most dramatic change in the interval between 0-10%.

Simple contagions create clusters immediately and spawn many clusters during the life-cycle. The minimally complex contagions ($T = 2$) show the most benefit with the first cluster appearing early in the contagion life-cycle and dozens of clusters emerging during the life-cycle of contagion on networks where long range ties are present. For complex contagions near the capacity of the network, few clusters appear and these clusters appear late in the contagion life-cycle when they are unable to accelerate the contagion process.



(a) Number of viable clusters formed.



(b) Proportion of nodes activated when first viable non-seed cluster forms.

Figure 2.7: Timing and frequency of cluster emergence for contagions of different thresholds on permutations of an 8-regular Moore lattice.

In Figure 2.6b the non-linear behavior after 50% long range ties is at least partially driven by network fragmentation. Before this point, all clusters eventually merge into a single large cluster. As graphs are permuted further, the short range ties are not sufficient to maintain a connected graph and the graph structure devolves into small components of nodes connected via short range ties with significantly disrupted neighborhood to neighborhood connectivity. These components are connected entirely via long range ties. Activation via long range ties is the only feasible contagion motif and this underlying structure contributes to the increased time to saturation and sharp cutoff of contagion success observed in Figure 2.5. This is consistent with the explanation for the sharp cutoff offered by Centola and Macy (2007).

The formation of clusters is subject to some stochastic variation because it depends on the occurrence of rare events. If the timing of cluster emergence is associated with contagion saturation time, then this relationship should also hold for multiple instances of contagion on networks of similar structure. Holding the proportion of long range ties constant, the emergence of clusters is a function of the threshold and stochastic variation in the location of the seed nodes and the update order. I grouped samples by the proportion of long range ties in the underlying network and contagion threshold and fit a simple regression predicting the iterations to saturation by the proportion of nodes active when the first viable cluster emerged (CL_p). CL_p was a significant predictor ($p < .0001$) for threshold 2 contagions on networks with between 0 and 40% long range ties. It was a significant predictor ($p < .05$) for threshold 1 and 3 contagions on networks with 0 to 10% long range ties. Within the ranges in which CL_p was significant, the proportion of adopters at cluster emergence was positively correlated with the number of iterations required to saturate the network.

Figure 2.8 reports the proportion of variation in time to saturation attributable to (CL_p). Among the thresholds considered, CL_p was the best predictor for threshold 2 contagions and only contributes to a small amount of the overall variation for the other thresholds. This is consistent with comparisons across networks with different proportions of long range ties, where the emergence of clusters was too late to benefit high threshold contagions. Simple contagions don't require multiple long range ties to establish clusters in remote parts of the graph, so the each node activated via a long range tie becomes remote cluster automatically. Figure 2.8 shows that the time a simple contagion first encounters a long range tie does impact the time to saturation, but only when ties are relatively rare. Once long range ties are common enough that a long range tie will be encountered within the first few adoptions, the timing of the encounter ceases to have much explanatory effect.

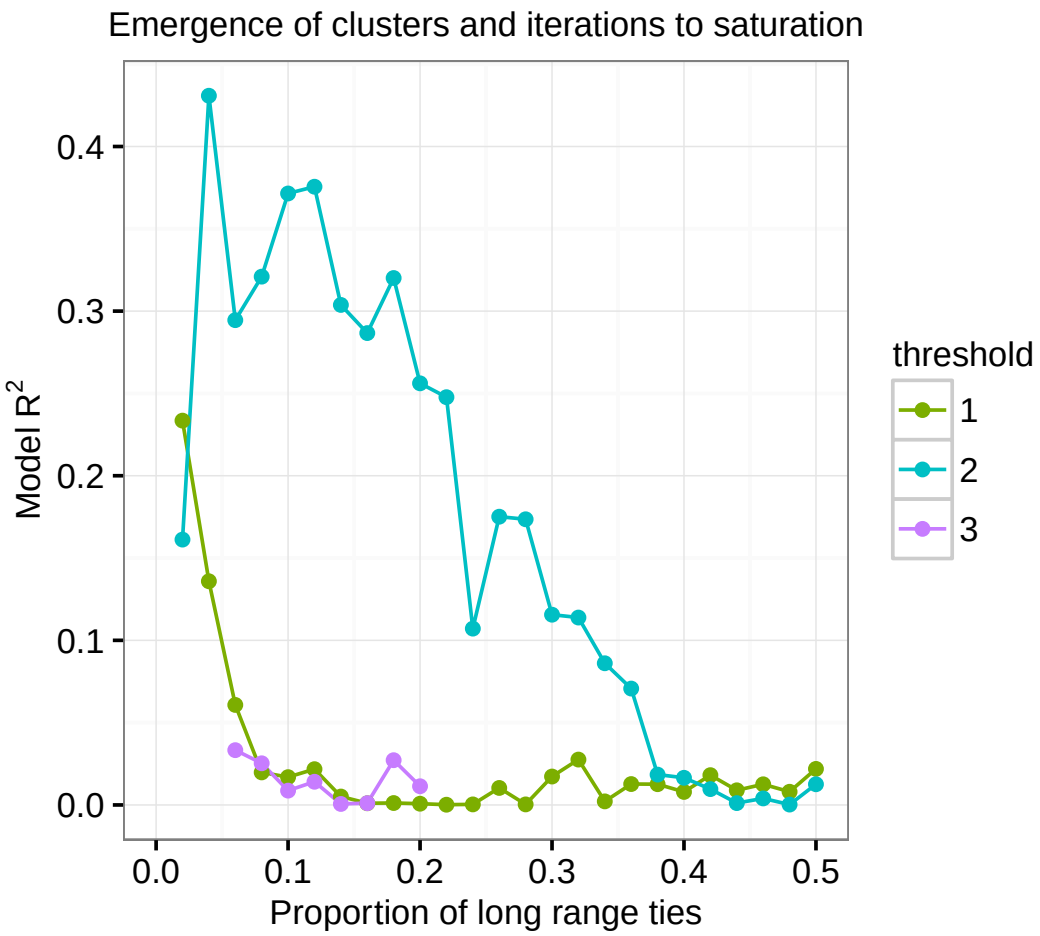


Figure 2.8: Proportion of variance in time to saturation associated with the time of first cluster emergence. Models fit at intervals for contagions of different thresholds on permutations of an 8-regular Moore lattice.

2.4 Discussion

The evidence presented in this chapter points to cluster emergence rather than contagion motif as the underlying mechanism through with the proportion of long range ties increases the adoption rate for simple and minimally complex contagions. The the number of nodes exposed via long range ties is relatively stable over the interval when time to saturation changes the most – the window around 10% long range ties. Most nodes are exposed via short range ties in

this interval. The distribution of activation via long range ties in the contagion life-cycle is not consistent with a mid-cycle increase in the probability of these events.

The changes in the timing of and frequency of cluster formation correspond to the decrease in time to saturation. Comparing across networks with different proportions of long range ties, only complex contagions below the transmission capacity of the network benefit from long range ties (Figure 2.5). For contagions near the capacity of the network ($T = 3$), long range ties form few clusters and do so late in the life-cycle. Simple contagions form new clusters easily and immediately as long as long range ties are present. Only complex contagions with thresholds below the transmission capacity can benefit from long range ties (Centola & Macy, 2007), and these contagions show the most significant changes in the number and onset of clusters as the number of long range ties increases (Figure 2.6b). The relationship between cluster emergence and time to saturation holds within network and across network.

CHAPTER 3

THE RELATIVE SHORTNESS OF LONG RANGE TIES

The preceding chapter identifies the emergence of clusters as a mechanism through which long range ties impact the adoption rates. Clusters increase contagion rates because they expose new nodes that are distance from the other regions of infection. Since clusters form from particular configurations of long range ties, the distribution of long range ties should impact the appearance of clusters. In particular, if long range ties are more likely between nodes that are close to each other, the network distance between the infected nodes and the nodes susceptible via long range ties would be tend to be smaller than on randomly permuted graphs. With this distribution of tie range, clusters would be more likely to appear near the periphery of the contagion perimeter, close to nodes already susceptible via short range ties. Crowded around the periphery of the main infected cluster, these new clusters would quickly merge with the seed cluster. Most new nodes exposed by the cluster would have been exposed to the seed cluster after a few waves of local spread. Since clusters reduce the time to saturation by increasing the number of nodes exposed to the contagion, these peripheral clusters would not be expected to reduce the saturation times. In this chapter, I consider the distribution of tie range in several empirical online social networks, compare the empirical distributions to those generated by the permutation algorithm used in prior complex contagion studies, and develop a permutation strategy that produces a more realistic tie range distribution while preserving other features of the Maslov—Sneppen permutation algorithm.

3.1 Previous findings

Though long range ties are an important element of sociological explanations (Grannovetter, 1973; Granovetter, 1983; Centola & Macy, 2007; Barash et al., 2012), the distribution of tie range has not been examined empirically. The computation of tie range has attracted significant attention of the past twenty years as the *second shortest path* problem in computer science but the literature focuses on the algorithm and not on empirical applications or surveys of social networks (Eppstein, 1994; Papaefthymiou, 1997; Li, Sun, & Chen, 2006; Kao, Chang, Wang, & Juan, 2011; Zhang & Nagamochi, 2012; Wu, 2013). A forthcoming work, Park (2016), aims to examine both the empirical distribution of tie range and possible mechanisms to produce the observed tie ranges in social networks. The present work is independent of Park (2016).

Most prior work examining the relationship between perturbed networks and contagion has relied on rewiring ties uniformly at random or creating new randomly directed ties (Watts & Strogatz, 1998; Centola et al., 2007; Centola & Macy, 2007; Barash et al., 2012). Kleinberg (2000) considered the distribution of long range ties in the context of decentralized search, introducing an alternative small world model. Recently, Ghasemiesfeh et al. (2013) considered contagion speed for stylized spatial networks with different distributions of weak ties. They consider a model where weak ties form at random – the Newman Watts model – and two models where the probability of weak tie formation depends on the proximity of the endpoints – the Kleinberg small world and hierarchical network models. Compared to contagions on networks with randomly formed long range ties, contagions on Kleinberg small world and hierarchical networks spread more slowly.

3.2 Tie range distribution in empirical social networks

To understand the empirical distribution of tie range, I selected an assortment of friendship and communication graphs drawn from online social networks, including networks from Flickr, LiveJournal, YouTube and Orkut from Mislove et al. (2007), the one hundred American college Facebook networks from Traud, Mucha, and Porter (2012) and Twitter data collected by the Social Dynamics Lab at Cornell University. These graphs range in size from thousands to millions of nodes and represent a variety of social networks organized around a variety of online activities. The computation of tie range is not a standard function of network analysis software so I implemented a custom tie range function using the iGraph C library (Csardi & Nepusz, 2006).

The proportion of long range ties in the surveyed networks ranges from .002 to .384. In general, larger networks have a larger proportion of long range ties but there is significant variation among networks of similar size. Table 3.1 and Table 3.2 report both the overall proportion of long range ties — P(LRT) — and the estimate k for the geometric distribution parameter for distribution of ties with range ≥ 3, which corresponds to the ratio between the number of ties at range x and the number of ties at range greater than x. For the purposes of this paper, it is sufficient to note that for the surveyed networks, most ties will be short range ties, most of long range ties will be length 3 and the frequency of ties with a particular range decreases dramatically as the range increases.

Table 3.1: Tie range statistics for large online social networks.

	Name	Nodes	Size	P(LRT)	k
Twitter:	Feb 2011	17,905	707,781	0.011	0.974
Egypt	Dec 2011	54,572	1,907,105	0.023	0.990
(all ties)	Jun 2012	94,140	3,029,650	0.038	0.988
	Aug 2013	109,082	3,425,747	0.081	0.981
Twitter:	Feb 2011	9,166	100,662	0.075	0.856
Egypt	Dec 2011	17,305	170,782	0.097	0.847
(mutual ties)	Jun 2012	28,805	211,671	0.144	0.758
	Aug 2013	41,879	235,483	0.205	0.629
Online	Orkut	2,997,354	113,910,527	0.134	0.914
Social	Flickr	1,192,171	13,779,153	0.054	0.774
Networks	LiveJournal	4,150,633	45,414,720	0.141	0.711
	YouTube	509,331	2,333,627	0.384	0.682

Table 3.2: Tie range summary for 100 Facebook networks.

Statistic	Nodes	Size	P(LRT)	k
Mean	12,070	469,838	0.017	0.967
St. Dev.	(9,067)	(363,540)	(0.011)	(0.016)
Min	762	16,651	0.002	0.920
Max	41,536	1,590,651	0.048	0.996

Overall, the surveyed networks have strikingly similar distributions of tie ranges — the pattern is qualitatively similar to a geometric distribution. Figures 3.1, 3.2, 3.3 show the proportion of ties with each range to illustrate the observed distribution. By definition, two is the minimum tie range for a graph without parallel edges. The ordinate axes is the logged proportion, which highlights the exponential decay in the number of ties at each range — the straighter the line, the more similar the distribution is to a geometric or negative binomial distribution. The empirical distributions are not a match for a geometric or negative binomial according to a goodness of fit test (Meyer, Zeileis, & Hornik, 2015); the number of ties with larger ranges tends to be larger than the expected

counts. Future work might explore the interaction of tie range and network diameter, which constrains the range distribution. In any case, the discrepancies from the geometric distribution are in the higher ranges and represent only a small fraction of the long range ties.

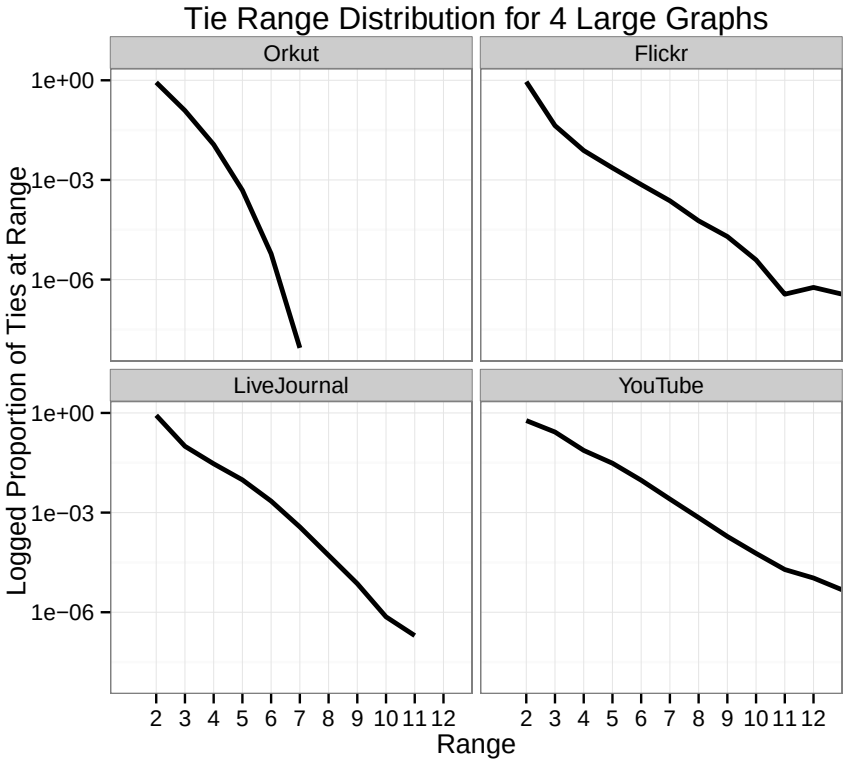


Figure 3.1: Percent of ties at each value of range for four large online social network graphs.

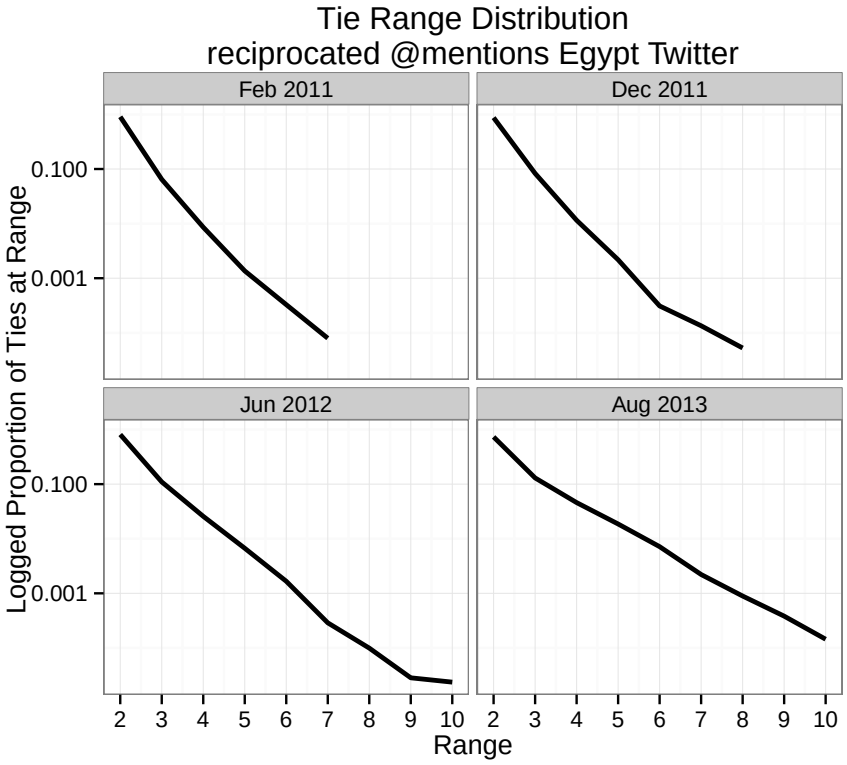


Figure 3.2: Percent of ties at each value of range in the reciprocated @mention network for Egyptian users active in four time periods.

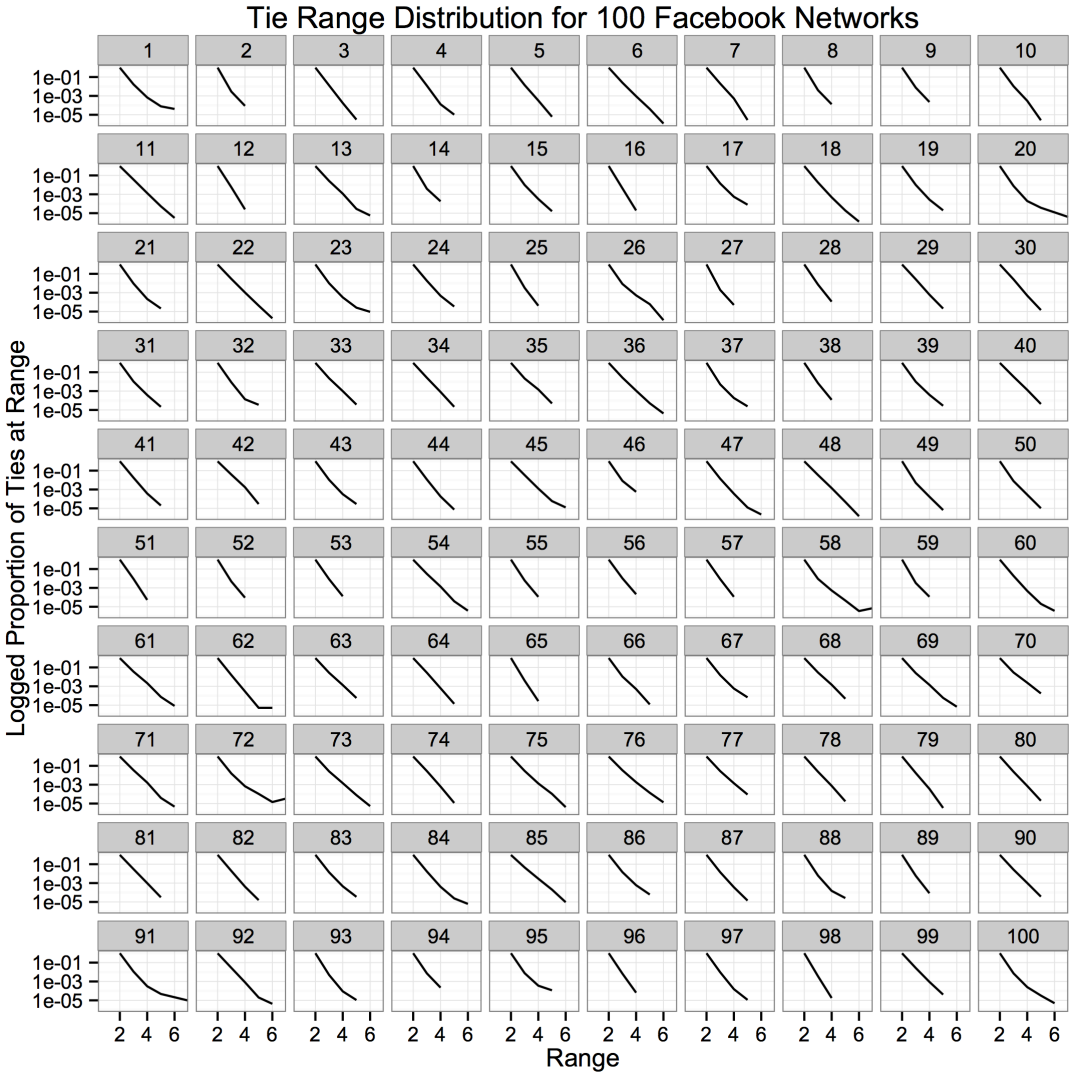


Figure 3.3: Percent of ties at each value of range in 100 Facebook friendship networks.

3.3 Tie range on MS perturbed lattice networks

Next, I examine the tie distribution that results from the perturbation algorithm introduced by Maslov and Sneppen (2002), sometimes called the double edge swap method. This approach has the benefit of preserving the degree of each node, thus preserving both the degree distribution and the density of the graph. Contagion processes are sensitive to both the shape of the degree distribution and the overall graph density, so Maslov—Sneppen (MS) rewiring can be used to examine the impact of tie perturbation while keeping other factors constant (Centola et al., 2007; Centola & Macy, 2007; Barash et al., 2012).

Figure 3.4 shows typical tie range distributions produced by MS rewiring at different levels of rewire. In the initial unperturbed lattice, all ties are range two. As these ties are randomly re-allocated to long range ties, the distribution of long range ties changes from flat to unimodal with a peak near the expected average path length, $\frac{\log N}{\log k}$ where N is number of nodes and k is the mean degree.

The convergence is apparent even at low levels of rewiring. This unimodal distribution is quite different from that observed in the empirical networks so it may be that the role of long range ties in the spread of contagions is different than generally recognized in the literature. In particular, if long range ties are not usually far-reaching, then the chance of spawning independent, locally-isolated clusters of infections becomes quite small. Instead, the contribution of long ranges ties will be focused near the boundary between infected and uninfected parts of the graph.

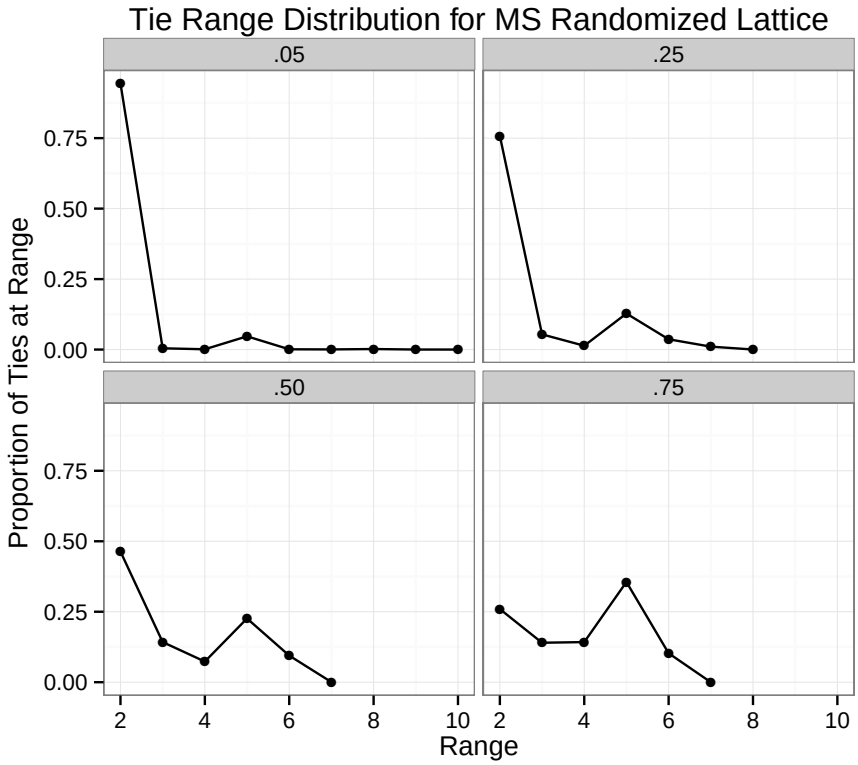


Figure 3.4: Percent of ties at each value of range for graphs with the proportion of long range ties equal to .05, .25, .50 and .75. These lattice graphs were rewired according to the Maslov–Sneppen method. Note that the axes are not logged.

3.4 Geometric biased double-edge-swap

A permutation strategy that is biased towards producing shorter long range ties would better match the empirical tie range distributions and could provide insight into how the complex contagion processes might unfold on more realistic networks.

Maslov—Sneppen rewiring (Maslov & Sneppen, 2002) operates by selecting two edges $A \to B$ and $C \to D$ uniformly at random and proposing two new edges by swapping the endpoints of the selected edges: $A \to D$ and $C \to B$. If either of the proposed edges already exists, then a new pair is selected at

random. If the proposed edges do not exist, they are added to the graph and the originally selected edges are removed. In this way, they network is rewired while preserving the degree of each node.

I created more realistic tie range distributions by modifying the way edges are selected for rewiring. Note that since all the lattice networks are degree regular and undirected, choosing a node at random and then a random edge incident to that node is the same as selecting an edge at random. The selection proceeds as follows:

1. Choose a source node at random.
2. Choose a distance d by taking a random draw from a geometric distribution with parameter = 0.95.
3. Choose a target node at random from the set of nodes with the shortest path from source node equal to d if such a node exists, otherwise restart.
4. Choose a one random edge incident to the source node
5. Choose a second random edge incident to the target node.
6. Apply the Maslov—Sneppen swap method to the two selected edges.

While not particularly efficient, this iterative approach generates tie range distributions with the same qualitative properties as the empirical distributions, particularly for the proportions of long range ties observed in the empirical networks. Figure 3.6 uses the same axes as the empirical networks and the downward trend is apparent up until the proportion of long range ties exceeds 0.50. This geometric biased rewire procedure (GEO) also creates distributions quite different from the the original Maslov—Sneppen permutation. Figure 3.5 uses

the same axes as Figure 3.4 to facilitation the comparison of the two permutation approaches. Most notably the geometric biased avoids the unimodal shape until there are a large fraction of long range ties and the maintains the scarcity of far reaching long range ties in relative to the shorter range ties.

The GEO permutation strategy produces graphs that are very similar to MS permutation in many respects. Both permutations preserve the number of edges and the degree distribution. Importantly, the edge overlap distribution is nearly the same in both graphs. Overlap is the number of neighbors the endpoints of an edge have in common – in other words, it is the width of the bridge between neighborhoods. Centola and Macy (2007) showed the existence of wide bridges was necessary for complex contagions to spread, so it is important that the permutation techniques manipulate overlap to a similar degree for the results to be comparable. Figure 3.7 shows the distribution of overlap for random (MS) and GEO rewired networks. The number of wide bridges between neighborhoods is very similar though slightly lower on GEO permuted lattices. Overall, any differences in the outcome of contagion processes are more attributable to large differences in tie range distribution rather than small differences in overlap.

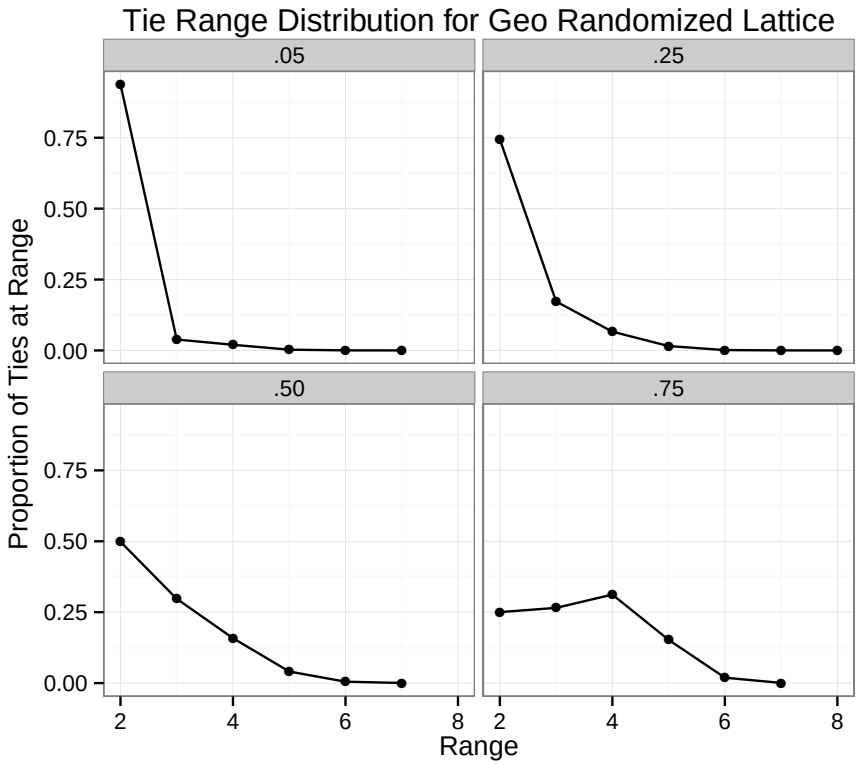


Figure 3.5: Percent of ties at each value of range for graphs with the proportion of long range ties equal to .05, .25, .50 and .75. These lattice graphs were rewired to favor the creation of shorter long range ties. Note that the axes are not logged.

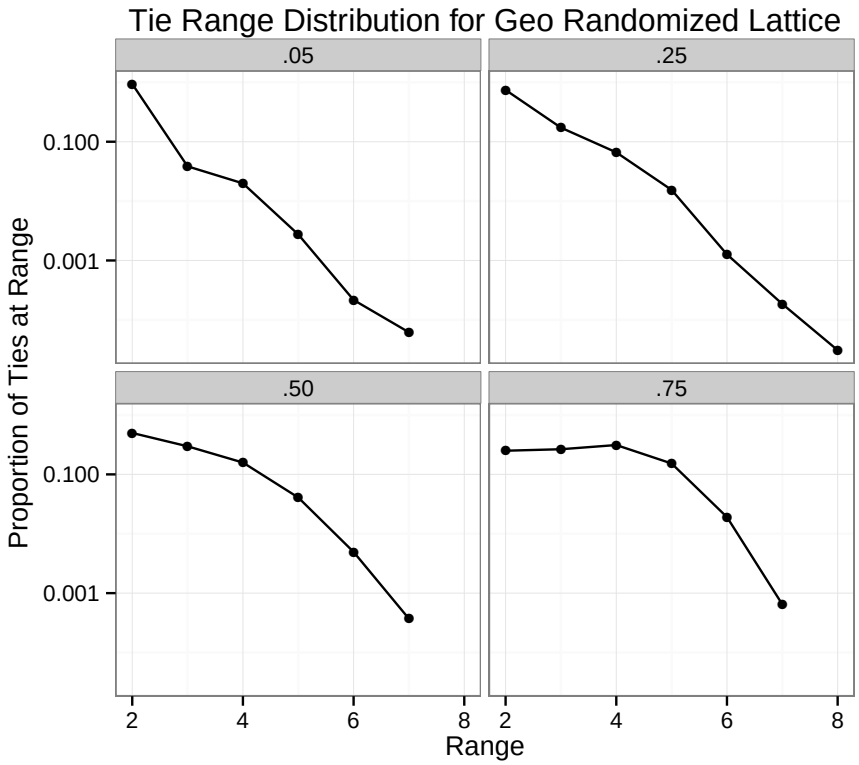


Figure 3.6: Percent of ties at each value of range for graphs with the proportion of long range ties equal to .05, .25, .50 and .75. These lattice graphs were rewired to favor the creation of shorter long range ties. The ordinate axis is logged to facilitate comparison with the empirical networks.

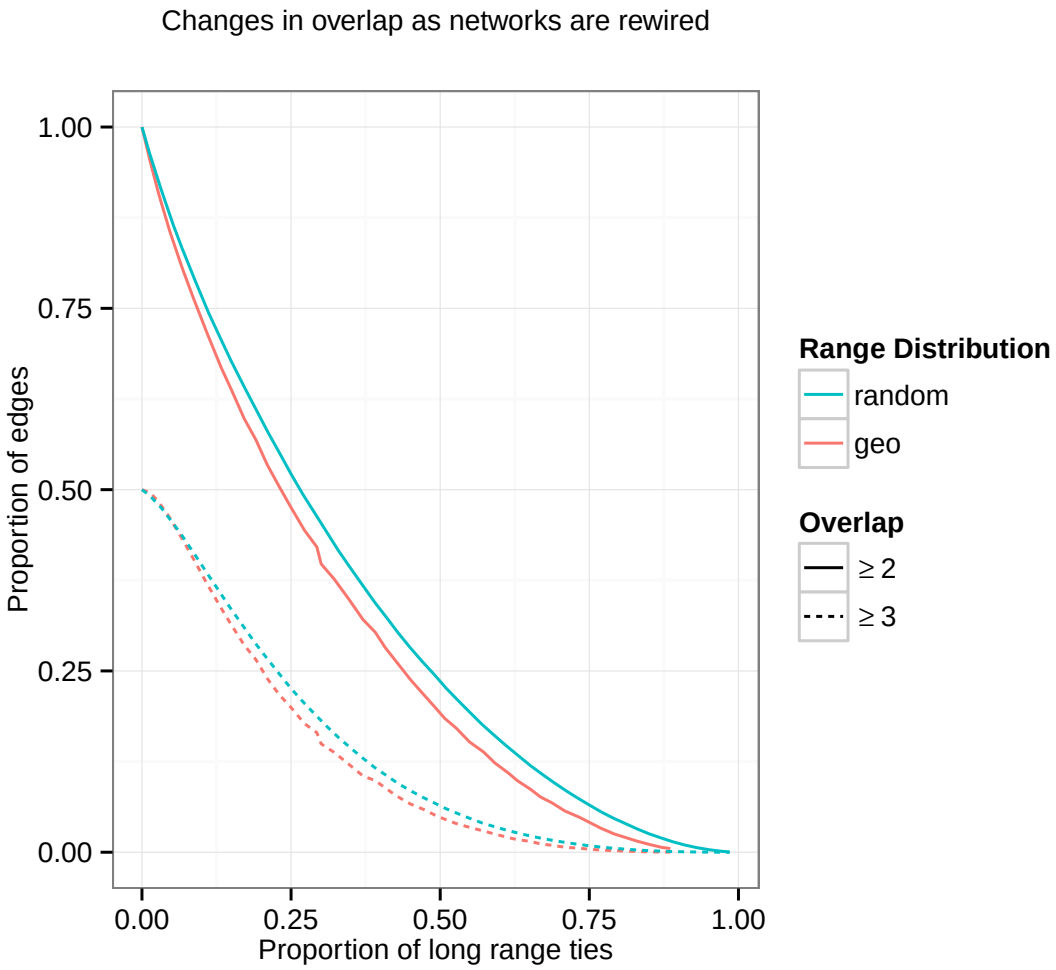


Figure 3.7: Comparison of neighborhood overlap under MS and GEO permuted 8-regular Moore lattice.

3.5 Discussion

A survey of several online social networks revealed that empirical tie range distributions are qualitatively similar in that most ties are range two and the preponderance of long range ties are range 3. The MS permutation used in previous simulations of complex contagion produces tie range distributions with fewer range three ties and more ties at the higher values of range than is typical of empirical networks. The analysis from Chapter 2 and a few examples from the literature show that the distribution of tie range can impact the rate a contagion spreads through the network so the discrepancy between MS and empirical networks could produce misleading model results. To support further investigation, I proposed the GEO permutation which produces a more realistic distribution of tie ranges while retaining the desirable properties of MS permutation.

CHAPTER 4

THE DISTRIBUTION OF TIE RANGE AND COMPLEX CONTAGION

In order to assess the impact of the distribution of tie range and relate to previous literature, especially Watts and Strogatz (1998) and Centola and Macy (2007), I used the previously described GEO permutation technique to generate a set of networks. GEO permutation prioritizes the creation of relatively short long range ties to generate a set of networks with tie range distribution more similar to that found in empirical networks.

I compare the spread of contagions with various thresholds over this set of networks to the spread of contagions on a set of networks that were perturbed using Maslov—Sneppen (MS) rewiring. The MS distributed long tie results are based on 200 perturbed networks with different levels of long range ties and 200 trials on each network. The GEO perturbed network tie results are based on 200 perturbed networks with different levels of long range ties and 200 trials on each network.

I first reexamine the relationship between the proportion of long range ties and contagion rate. I then describe the spatial patterns and adoption curves for representative instances of contagion finally present summary measures for indicators of contagion dynamics for multiple contagion processes on networks with a range of proportions of long range ties. I note major differences between the contagion processes on MS and GEO permuted lattices.

4.1 The distribution of ties and the rate of contagion

For simple contagions, random ties decrease the time for the contagion to saturate a network (Watts & Strogatz, 1998). Centola and Macy (2007) demonstrated that randomly rewiring a network has a non-monotonic effect on the number of time steps required for a complex contagion to saturate a network; low levels of rewiring provide no advantage, moderate levels of rewiring decrease the time to saturation and higher levels of rewiring actually increase time to saturation. Centola and Macy (2007) also identified the existence of a critical upper limit for the level of rewiring for complex contagions. Though the particular value of the critical point depends on the contagion threshold, all complex contagions are unable to spread if a network is sufficiently randomly rewired.

Prior work in Centola and Macy (2007) and Barash et al. (2012) characterized the amount of perturbation in terms of the proportion of rewired ties which was used as a proxy for the presence of long range ties. Unfortunately, the proportion of rewired ties does not directly assess the proportion of long range ties and it is hard to express the configuration of an empirical network in terms of the proportion of rewired ties. While both these papers provide important theoretical insight into the dynamics of complex contagion it is hard to relate their main findings to processes on empirical networks without establishing their relationship to the proportion of rewired ties. I reproduce some results from Centola and Macy (2007) but reframed them in terms of the proportion of long range ties.

The major elements of Centola and Macy (2007) and Watts and Strogatz (1998) are still apparent when re-framed in terms of the proportion of long range

ties. The results for MS perturbed networks are shown in Figure 4.1. When the threshold is zero, a random node is updated at each time step, so this serves as a baseline for the number of time steps necessary to update all the nodes. For the simple contagion — when the threshold is one — even a small fraction of long range ties decreases the time to saturation. For the minimally complex contagion with threshold two, creating long range ties initially decreases the time to saturation. Like simple contagions, minimally complex contagions reach saturation faster in the presence of some long range ties. In contrast, minimally complex contagions require a higher fraction of long range ties to reduce the saturation times. At higher fractions of long range ties, the loss of local short range ties begins to slow the spread of the contagion through the network, eventually reaching a critical point where the lack of dense local structure prevents the spread of the contagions. On a Moore lattice graph with degree 8, the threshold three contagion is the maximum threshold that can propagate. This high threshold contagion depends on the dense local structure to spread and any reallocation of short range ties to long range ties is detrimental. The time steps to saturation for the high threshold contagion increase rapidly with the proportion of long range ties, quickly reaching the critical point.

Repeating the same simulation on otherwise identical networks permuted with the geometric random edge swap method reveals some interesting consequences of the distribution of tie range — overall, the effect of long range ties is dampened. Figure 4.2 summarizes the simulation results for GEO perturbed nets.

When most long range ties are relatively short ranged, the simple contagion benefits less from long ties. As with MS permutation, additional long range ties

are never detrimental for simple contagions.

The spread of complex contagions on networks with different tie range distributions differs in a few notable respects. For the minimally complex contagion on a GEO permuted network, the long range ties initially decrease the time to saturation. This is similar to the results for MS networks, but a higher proportion of long range ties are necessary to achieve the same decrease in saturation times. The detrimental effects of the loss of local structure on threshold 3 contagion are less pronounced, suggesting that the shorter-range long ties somewhat offset the loss of short ties. GEO networks still have a critical permutation threshold beyond which a complex contagion can not spread but the critical point is at a higher proportion of long range ties.

The second pronounced difference between the MS and GEO results is the lack of a marked increase in the time to saturation for the values slightly less than the critical proportion of long range ties. This could indicate that the transition is very abrupt or that the underlying graph structure that prevents contagions from spreading is not the same as the mechanism that contributes to increased time to saturation. One possible explanation is that the degraded local structure at high levels of rewiring result in few viable seed clusters. Compared to MS permuted networks, GEO permutation results in a distribution of tie range where long range ties are both less helpful at lower proportions of long range ties and less detrimental to the spread of contagions at higher proportions of long range ties. One possible explanation is that the degraded local structure at high levels of rewiring result in few viable seed clusters.

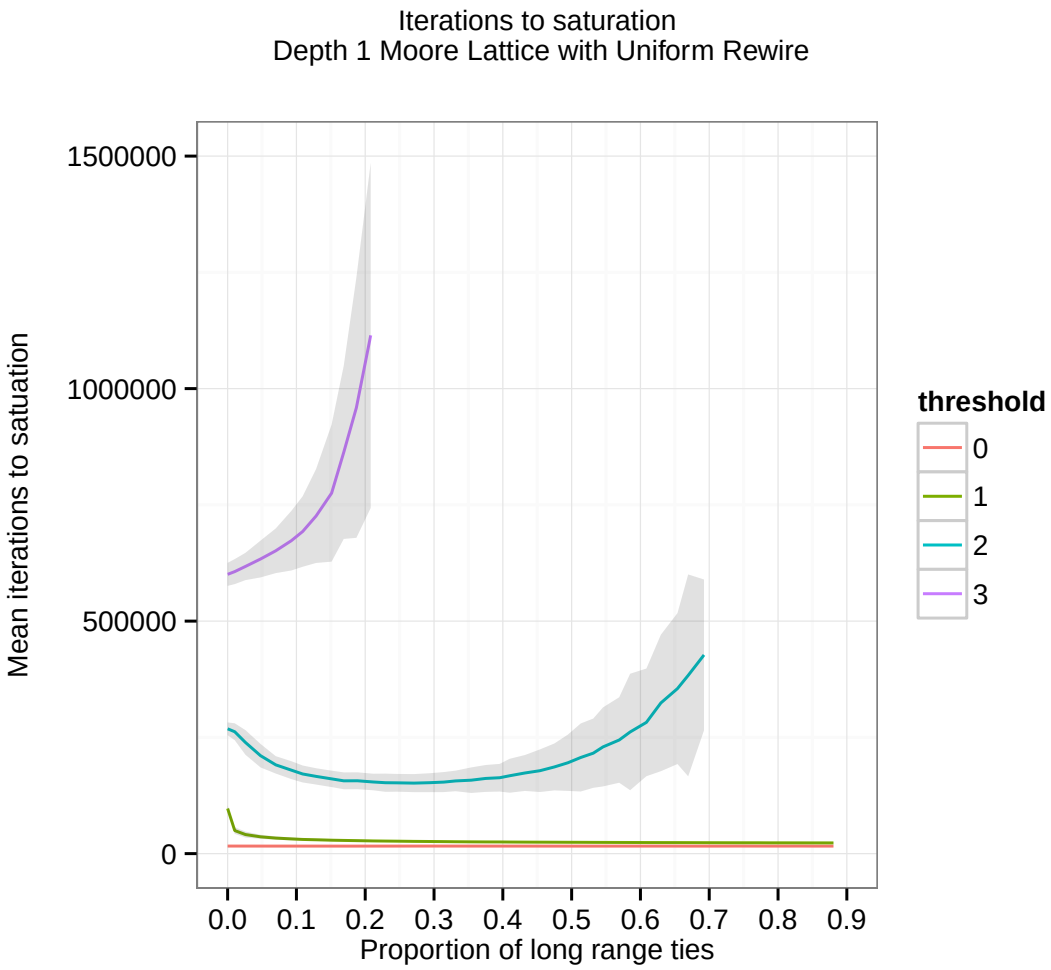


Figure 4.1: Mean iterations to saturation for contagions with various thresholds on a MS permuted 8-regular Moore lattice.

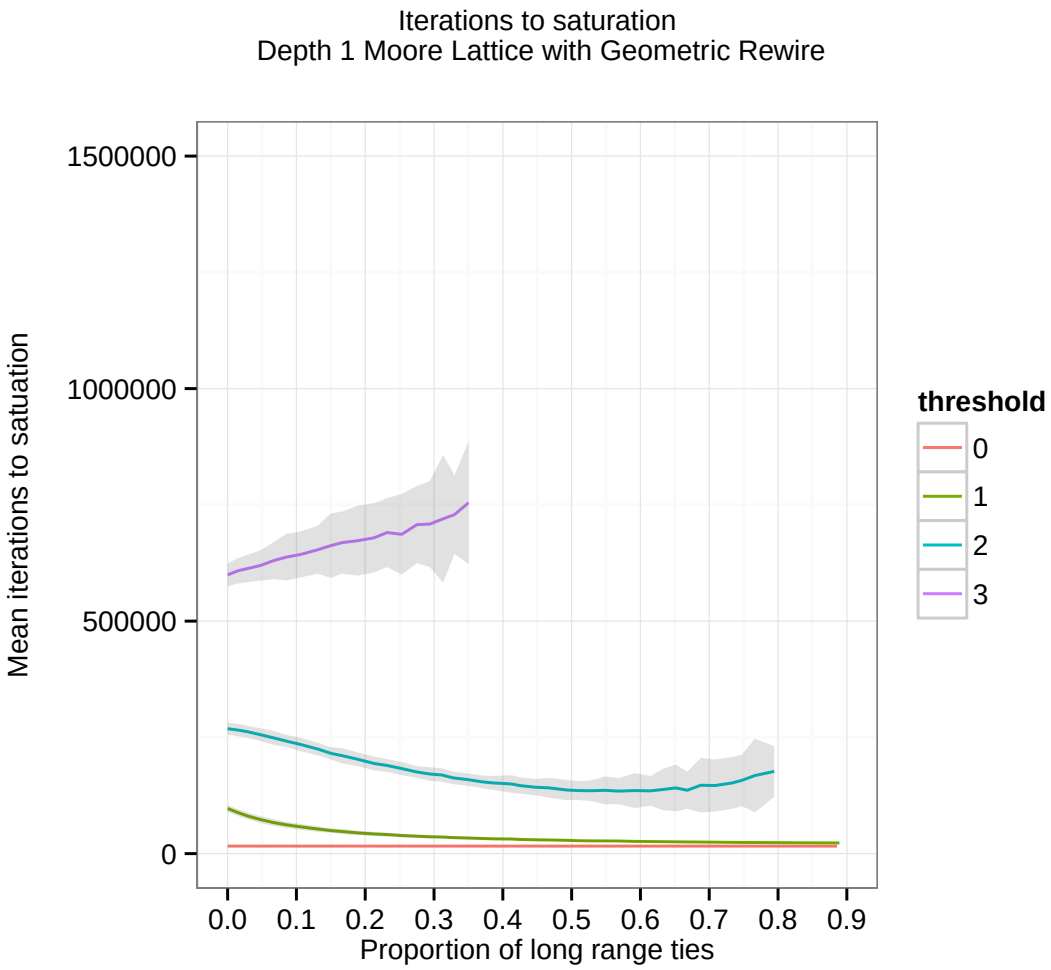


Figure 4.2: Mean iterations to saturation for contagions with various thresholds on a GEO permuted 8-regular Moore lattice.

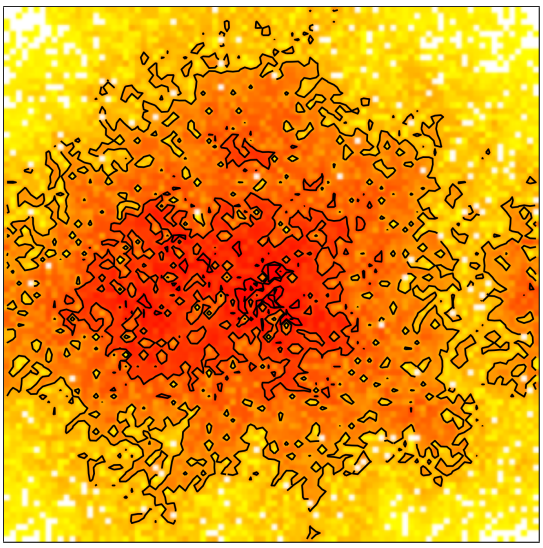
4.2 Contagion dynamics

Following the structure of the discussion of contagion on MS permuted lattices, this section explores the spatial patterns and adoption curves for representative instances of contagion, highlighting and discussing differences between the contagion processes on MS and GEO permuted lattices.

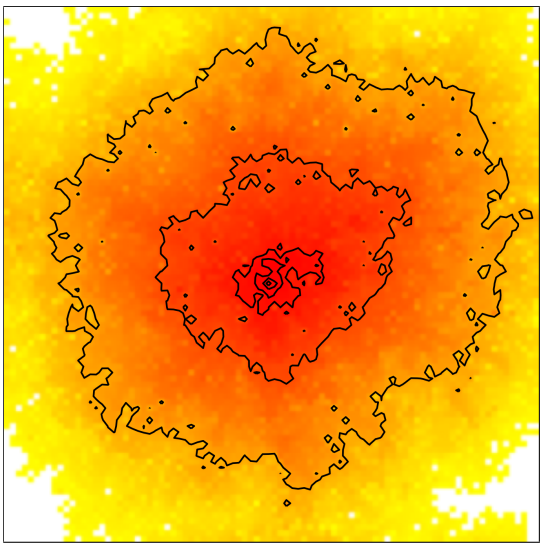
4.2.1 Contagion motifs and phases

Figure 4.3 shows the dynamics of contagion on a GEO permuted 8-regular lattice of 8100 nodes with 10% long range ties. In each of the sub-figures, the network structure and seed cluster are exactly the same - the only difference is the contagion threshold. For ease of comparison, the seed cluster is centered in each image. The contour lines show the extent of adopters at 20, 100, 1000 and 4000 nodes adopted. In comparison to the results for MS permuted lattices reported in Figure 2.2, contagion on the GEO permuted lattices is more centered around the seed cluster for both simple and complex contagions.

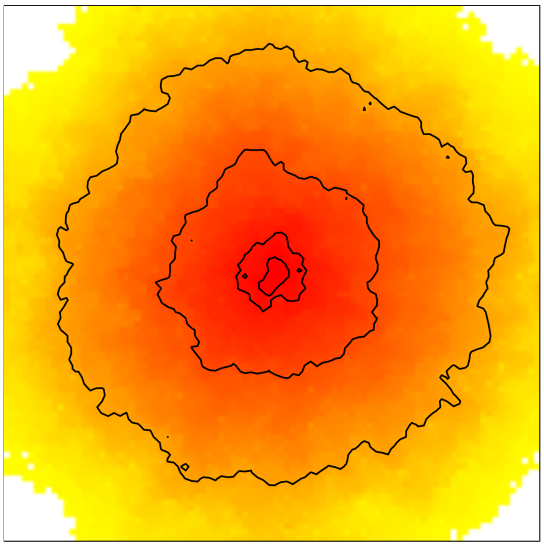
For a simple contagion, (Figure 4.3a) the long range ties only rarely reach nodes distant from the seed cluster so the radial pattern more typical of complex contagions is apparent even for simple contagions on GEO permuted networks. Most nodes activated by long range ties are relatively near other adopters, creating a ragged boundary between adopters and non-adopters. The ragged boundary exposes more nodes than a smooth boundary, so simple contagions still benefit from more long range ties. Figure 4.2 shows that mean time to saturation decreases for simple contagion as the proportion of long range ties increases.



(a) Threshold = 1



(b) Threshold = 2



(c) Threshold = 3

Figure 4.3: The history of contagion on a spatial representation of a GEO permuted lattice graph. Pixels represent individual nodes at corresponding coordinates lattice coordinates and pixel color reflects order of activation. Red nodes activated first with lighter colors activating last.

Comparison with figure 4.1 shows that randomly distributed long range ties generated by MS permutation are more beneficial than a similar number of GEO distributed long range ties for simple contagions ($T = 1$).

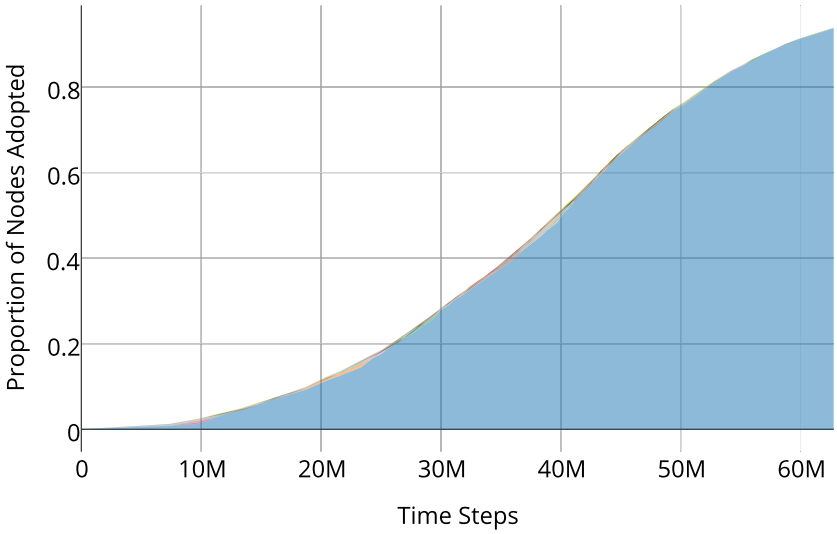
For the complex contagions the radial spread from the initial seed cluster is pronounced. The lower threshold complex contagion shows no indication of cluster emergence (Figure 4.3b). The long range ties do create some raggedness around the boundaries between adopters and non-adopters but the the overall pattern is neighborhood to neighborhood spread radiating from the central seed cluster. Comparing Figures 4.2 and 4.1, a larger proportion of GEO distributed long range ties are required to reduce the contagion saturation times. Comparing the behavior for threshold 2 contagions on the right hand side of these figures, it is apparent that for higher networks with a higher proportion of long range ties, GEO permutation is less detrimental than MS permutation. The distribution of tie range does not change the contagion pattern for contagions with thresholds near the capacity of the network (Figure 4.3c) because these contagions rarely establish viable clusters of adopters or do so late in the life-cycle after many nodes have already adopted. Constrained to spread from via wide bridges, these contagions radiate from the seed cluster via neighborhood to neighborhood spread. A higher proportion of long range ties is never beneficial to higher threshold contagions, but a comparison between Figure 4.2 and Figure 4.1 shows that GEO distributed long range ties actually support higher threshold complex contagion better than randomly directed long range ties. High threshold complex contagions spread relatively faster and tolerate a larger proportion of long range ties on GEO permuted networks. In GEO permuted networks where randomization has replaced some wide neighborhood to neighborhood bridges with long range ties, the long range ties are more likely to be directed at nearby neighborhoods, creating wide bridges between infected and uninfected nodes. Compared to MS permutation, GEO permutation partially offsets the loss of neighborhood to neighborhood bridges with an

increased probability of cluster to neighborhood ties.

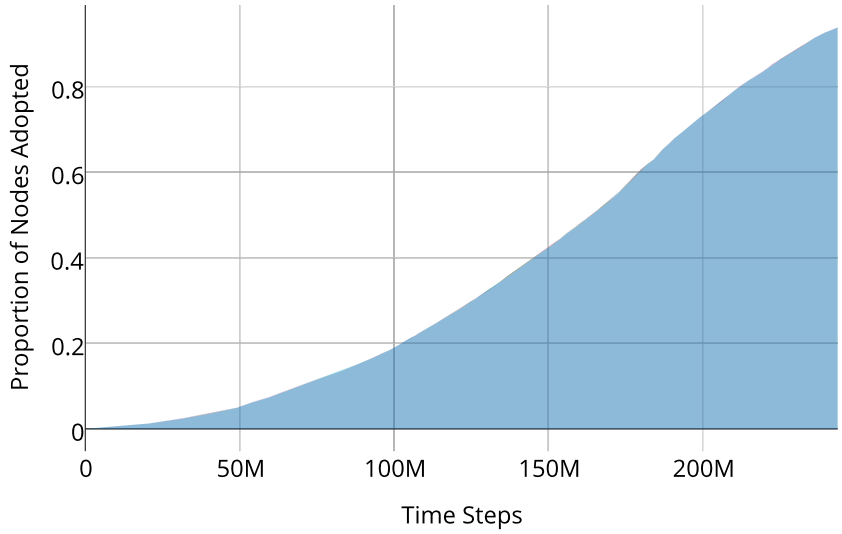
4.2.2 Life-cycle differences in simple and complex contagion

The adoption curves on GEO permuted lattices do not show signs of significant isolated cluster development for simple or complex contagion. Scattered small clusters develop during the life cycle of the simple contagion shown in Figure 4.4a. In contrast to the numerous locally isolated clusters that develop on MS permuted lattices (Figure 2.3a) the clusters on GEO permuted lattices are quickly absorbed into the main cluster and there is not a pronounced acceleration in the adoption rate.

For complex contagions (Figure 4.4b), no clusters are apparent. The growth rate accelerates very slowly and the sudden increase in the adoption rate that corresponds to the appearance of clusters in contagions on MS permuted networks never occurs (c.f. Figure 4.4b).



(a) Threshold = 1



(b) Threshold = 2

Figure 4.4: Adoption of a contagion on a GEO permuted lattice over time with clusters highlighted. The infected nodes are grouped into locally isolated clusters. The width of each colored band shows the fraction of nodes within that cluster at that time step.

4.3 Long range ties and complex contagion

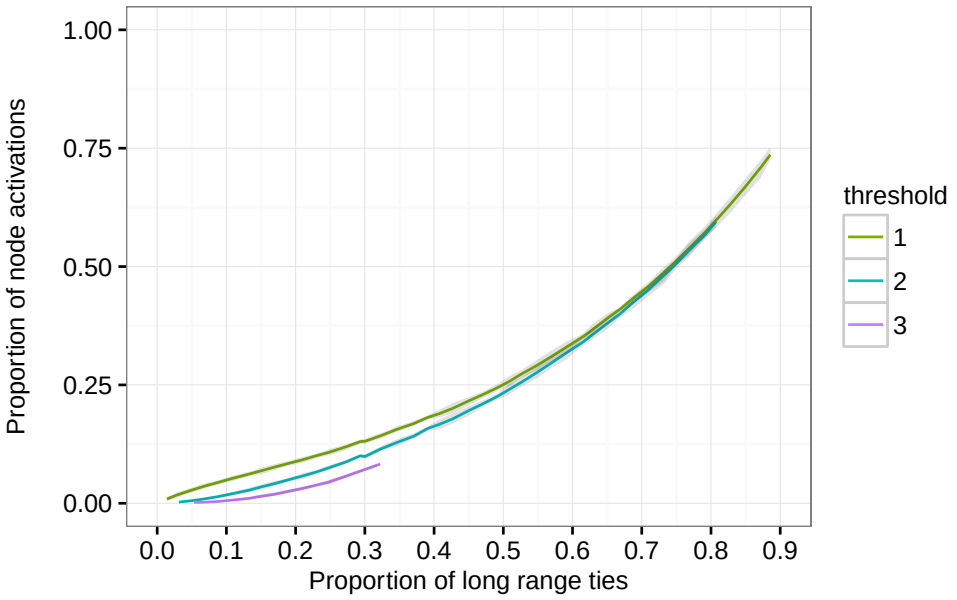
Having examined specific cases of contagion in the previous section, now I present summary measures for indicators of contagion dynamics for multiple contagion processes on networks with a range of proportions of long range ties.

4.3.1 Contagion motif dynamics

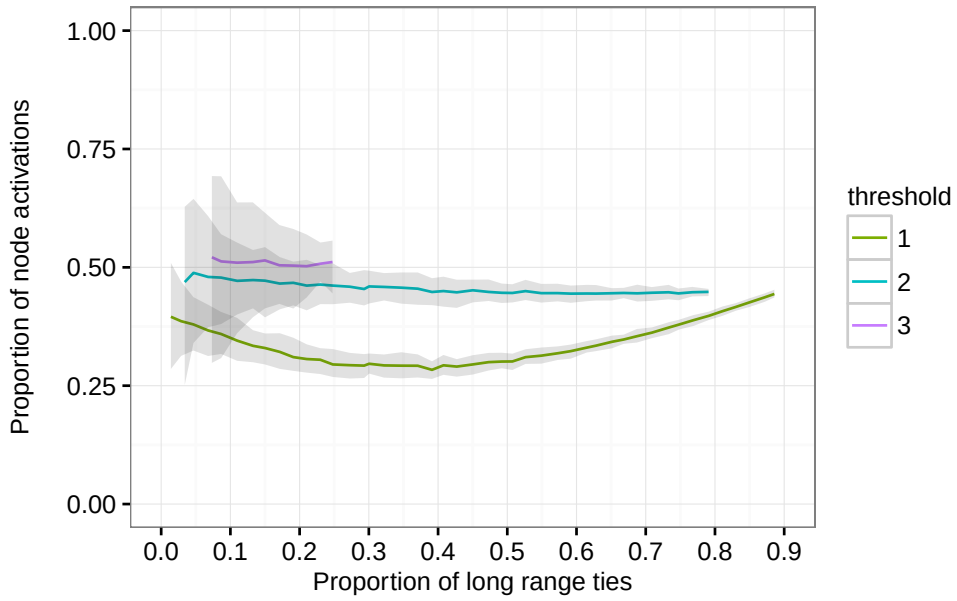
Unsurprisingly, the fraction of node adoptions attributable entirely to long range ties increases with the proportion of long range ties in the graph. Figure 4.5a shows the proportion of node activations attributable entirely to long range ties on GEO permuted lattices and the pattern is nearly identical to that observed on MS permuted lattices (Figure 2.6a). Again, considering the window around 10% long range ties, it is clear that local spread—perhaps aided by some long range ties—remains the dominant contagion motif. Similar comparisons of the proportion of nodes activated exclusively via short range ties and the proportion of nodes activated by a combination of short and long range ties reveal no discernible difference in networks permuted with different strategies.

Though the number of activations by long range ties is comparable, the time distribution of these activations is different on GEO permuted networks compared to MS permuted networks. On both types of networks, activations via long range ties occur throughout the contagion life-cycle, but the median time of activation tends to be closer to the median node activation time in GEO networks. Because long range ties tend to be shorter in GEO permuted networks, the probability that a random node will be activated by long range ties is less

sensitive to the number of active nodes in the network. For complex contagions that benefit from long range ties ($T = 2$), ninety percent of the activations by long range ties occur between 5 and 90% of all node activations with about equal density in the first and second half of the contagion life-cycle. In GEO networks a node's probability of activation via long range ties is more sensitive to its proximity to the infected cluster. In some ways, infection via long range ties is easier on GEO networks and this results in activation via long range ties earlier in the contagion life-cycle.



(a) Proportion of nodes activated via long range ties.



(b) Proportion of nodes active at median long range activation time.

Figure 4.5: Timing and frequency of activation via long range ties for contagions of different thresholds on GEO permutations of an 8-regular Moore lattice.

4.3.2 Cluster dynamics

Though node activation via long range ties occurs slightly earlier in the contagion life-cycle, this does not result in more frequent or earlier formation of locally isolated clusters on GEO permuted networks. The proportion of long range ties still has a marked impact on the formation of clusters but fewer clusters reach a viable size while isolated. In contrast to MS permuted networks, even viable clusters ($n \geq 9$) rarely grow to any significant size before merging with the original seed cluster.

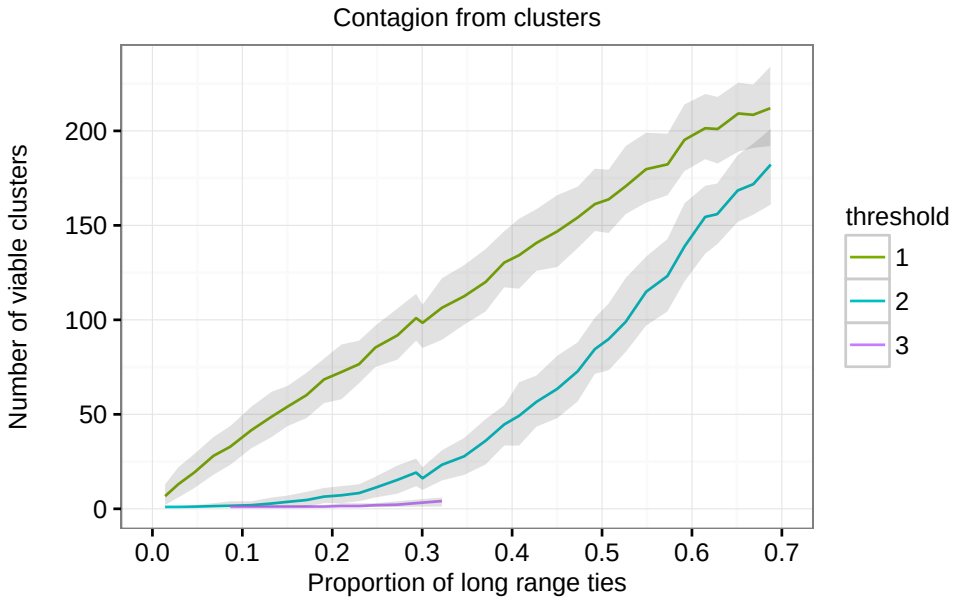
Compared to the contagion process on MS permuted networks, the relationship between the proportion of long range ties and the number of viable clusters produced in the life-cycle is more complicated on GEO permuted latices. For simple contagions, the number of clusters increases linearly with the proportion of long range ties in the network. For complex contagions very few clusters emerge at lower proportions of long range ties (Figure 4.6a). Minimally complex contagions eventually generate clusters on networks with a higher proportion of long range ties. The high threshold contagions ($T = 3$) never generate an appreciable number of clusters.

Non-seed cluster emergence is later in the life-cycle on average for contagions on GEO permuted networks but also much more variable (Figure 4.6). Because most clusters will be close to the original seed cluster, many clusters will not be able to reach a viable size before encountering and merging with the seed cluster.

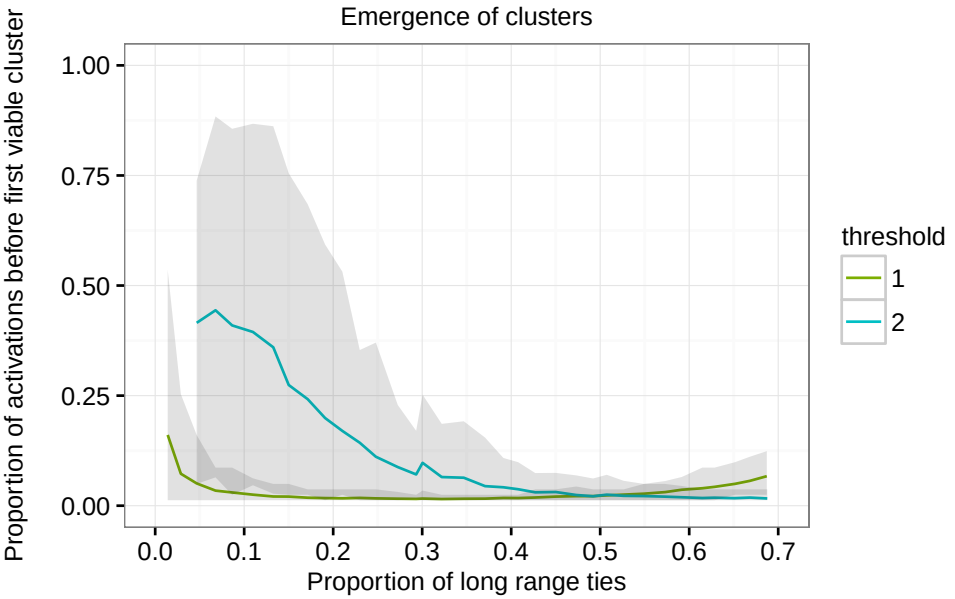
For GEO networks, I grouped samples by the proportion of long range ties in the underlying network and contagion threshold and fit a simple regression pre-

dicting the iterations to saturation by the proportion of nodes active when the first viable cluster emerged (CL_p). It was a significant predictor at the ($p < .05$) level for simple contagions on networks with less than 50% long range ties. Within the ranges in which CL_p was significant, the proportion of adopters at cluster emergence was positively correlated with the number of iterations required to saturate the network. The proportion of variance explained by these models was small with a mean r-square of 0.06 and a max of 0.12. CL_p was no longer a significant predictor at the ($p = .05$) for threshold 2 contagions and clusters did not emerge with enough frequency to estimate the model for threshold 3 contagions.

Cluster emergence is a less important predictor for GEO permuted networks because the clusters that do form rarely expose nodes in a distant part of the graph. Though they initially expose new nodes, these clusters soon merge with the original seed cluster and their influence is short lived. Figure 4.7 shows the size of the largest non-seed cluster for GEO and MS permuted graphs with different proportions of long range ties. The average size of the largest clusters for contagions in MS permuted graphs regularly account for more than 10% of the graph size. In contrast, the average size of the largest non-seed cluster is just a few percent of the total graph size, especially in the interval around 10% long range ties. For a given proportion of long range ties, a complex contagion process will generate fewer clusters and shorter-lived clusters later in the contagion life-cycle. On GEO networks, the emergence for the first cluster is a poor predictor of time to saturation for complex contagions because clusters rarely emerge and many clusters that do emerge are short lived.



(a) Number of viable clusters formed.



(b) Proportion of nodes activated when first viable non-seed cluster forms.

Figure 4.6: Timing and frequency of cluster emergence for contagions of different thresholds on GEO permutations of an 8-regular Moore lattice.

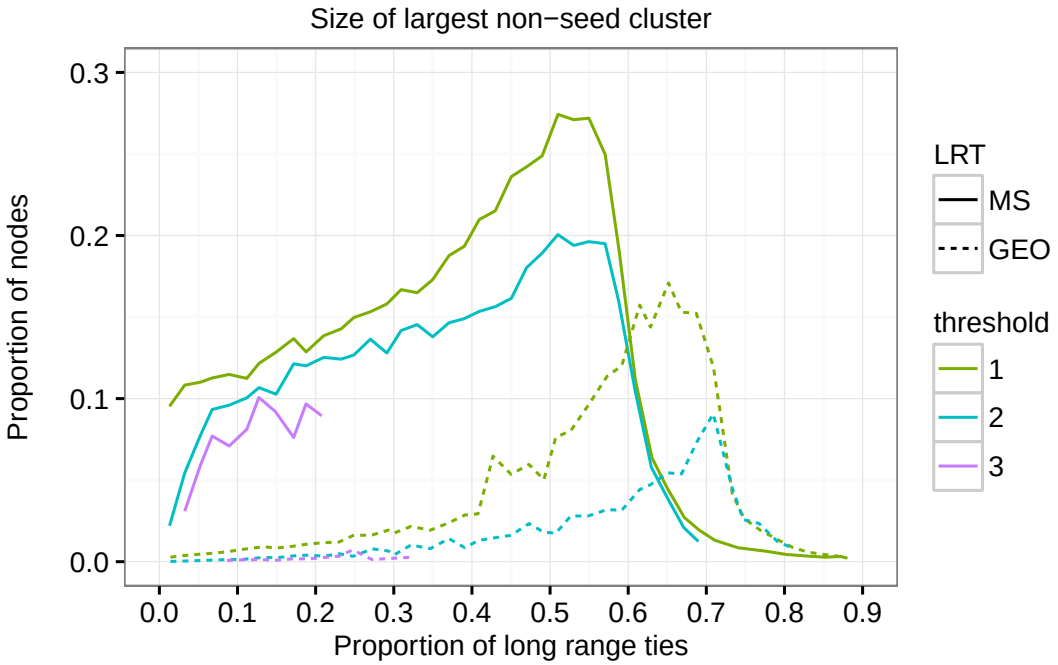


Figure 4.7: Proportion of nodes in largest non-seed cluster before merging for contagions of different thresholds on GEO and MS permuted networks

4.4 Discussion

The overall effect of GEO permuted long range ties compared to MS permuted ties is to dampen the impact of long range ties. Compared to MS permutation, it takes more long range ties to increase the rate of propagation and to also more to cause a detrimental effect.

Instead of clusters, contagions in GEO permuted lattices form deep perimeters, a new kind of contagion pattern, where long range ties activate nodes a little beyond the boundary between adopters and non-adopters and then expose the intermediate nodes from both sides. The pattern of spread on GEO permuted lattices is markedly more radial than on MS lattices, even for simple contagions and the role of clusters is significantly reduced. Even minimally complex contagions are unable to establish viable clusters. The adoption curves show a pattern of smooth acceleration without identifiable phases for all levels of threshold.

Despite the differences, the contagion motifs are nearly the same – in both types of networks, activation via long range ties is rare in networks with realistic proportions of long range ties. The distribution of long range ties particularly impacts the probability that the network structure will support the endogenous formation of locally isolated clusters. If another mechanism were to produce locally isolated clusters in a GEO network, the cluster would result in the same increase in exposure. It may be that empirical networks do not have a structure that can produce sudden acceleration of the adoption rates endogenously, but they may still be susceptible to rapid acceleration by the action of instigators or events that create small clusters of new adopters.

CHAPTER 5

ESTIMATING THRESHOLDS FROM OBSERVATIONAL DATA

5.1 Introduction

The preceding chapters emphasize the differences in the course of the contagion process among contagions of different thresholds. When investigating a suspected empirical contagion, determining the threshold of the contagion is one of the most informative measures about the future course of the contagion process. Unfortunately the threshold is difficult to measure accurately using observational data.

In a more realistic model of behavioral contagion on a network, an individual *ego* adopts a behavior if at least T of their *alters* adopt the behavior. For a given ego and behavior, T is ego's activation threshold. Unlike the models in previous chapters, the threshold varies from person to person, so there is a distribution of thresholds in the population.

Complex contagion models assume the existence of behavioral contagions for which many nodes have an activation threshold greater than one. The theory predicts that the dynamics of simple and complex contagion will differ so a reasonable task would be to distinguish empirical behavioral contagions with thresholds generally greater than one from those with thresholds generally less than on equal to one.

If the observed exposure to already adopted neighbors at a node's adoption time can be linked to a node's unobserved internal threshold, then observational data can be used to classify an empirical behavioral contagion as simple or com-

plex and to predict the underlying distribution of thresholds in the population. Cosley, Huttenlocher, Kleinberg, Lan, and Suri (2010) formalized a method for computing the proportion of nodes that adopt a contagion after a particular level of exposure. The result of this is computation is a $p(k)$ curve that gives the observed probability that a node with exposure equal to k will activate before thier k + 1-th exposure. Romero et al. (2011) computed the $p(k)$ curves for different categories of hashtags and found that there were notable difference in the probability of adoption at higher levels of exposure. In particular, high levels of exposure to political hashtags continued to be associated with relatively high levels of adoption while the impact of multiple exposure declined for more conversational tags. If political hashtag use is more risky than using conversational hastags, these results appear consistent with the complex contagion model.

The implicit link between observed exposure and activation threshold requires deeper investigation. Valente (1995, p. 75) discusses a observed threshold lag between the triggering exposure and the adoption time but treats the personal network status as constant during this lag time. In this framework, the individual's decision might not be immediately apparent in their behavior due to time spent planning or gathering resources. The possibility that network exposure might change during the lag is acknowledged in a footnote. I argue that the interrelated states of nodes in the network virtually guarantees that observed exposure overestimates true adoption thresholds and propose a measurement technique that improves estimates of node thresholds from observational data.

5.1.1 A problem of interrelated states

Consider a hypothetical laboratory experiment designed to measure a subject's threshold for adopting some particular contagion. In this experiment, an individual and his peer group enter the laboratory and the focal subject observes his peer group adopting a contagion. Unknown to the subject, the peer adoption is controlled by the experimenter. Peers adopt sequentially, one per round, and the subject is monitored for adoption after each activation round. Assume adoption is instant once the threshold is reached. In this experimental setting, the subject's adoption status is measured at each level of exposure, so the the subject's threshold can be determined with high accuracy. The experiment is costly because it requires using a whole ego network to gather a single observation.

In an alternative experiment, a dice roll determines how many peers adopt each round. In this experiment, the measure of the subject's threshold can be determined to be within an interval greater than their exposure during the round before they adopt and less than or equal to their exposure at the time of adoption. The true threshold is always less than or equal to the maximum value in the interval, so measuring the subject's exposure at activation results in an estimate greater than or equal to the true value.

On a platform like Twitter, where users are not constantly logged in, observational data corresponds more to the second experiment. Users check their account at intervals, so their exposure can jump several levels without them ever having experienced the intervening levels of exposure. The size of the threshold interval is limited only by the number of network neighbors.

Consider a fully connected graph of size m where nodes update their status in a random order. If every node has a threshold of 1 and one node in the cluster is activated, then we will observe activation exposures of 1, 2, ..., m-2, m-1 as each successive node in the cluster activates. Only one of the nodes in the cluster will have an activation exposure equal to its threshold and the rest will have activation exposure greater than their true threshold.

The interdependent state of the nodes presents a challenge if a researcher needs a $p(k)$ curve that corresponds to a hypothetical experimental outcome but only has observational data to use. This is particularly the case for research that relies on logs of user activity collected from online social networks.

5.1.2 Computing $p(k)$ curves

The $p(k)$ computation in Cosley et al. (2010) and Romero et al. (2011) is the observed conditional probability of activation with exactly k-exposure given non-activation at k-exposure. It is alternatively described as the fraction of nodes who were ever k-exposed who activated before their k + 1-th exposure. For nodes that activate, exposure is measured at activation time. For nodes that never activate, exposure is measured at the end of the observation window. Assuming that nodes constantly monitor their exposure and node activation is asynchronous, then this conditional probability is computed as

$$p(k) = \frac{\text{all nodes that activate with exposure } = k}{\text{all nodes with exposure } \geq k} \tag{5.1}$$

This computation makes the critical assumption that if a node has k expo-

sure, then it must also have had exposure equal to each value in $\{1, 2 \ldots, k-1\}$ at some point in the past. The $p(k)$ curve can be interpreted as the observed probability of activation before $k+1$ exposures for a random node with exposure equal to k at a random time in the observation window.

I also consider an alternative formulation where node activations in a neighborhood are assumed to happen nearly simultaneously so that once a node's exposure is greater than 0, its exposure does not change significantly before the node itself activates. Under this model, the k-exposed nodes were never exposed, at the intermediate levels, but went directly from 0 to k exposures. Therefore the nodes with exposure greater than k do not contribute information about the probability of adopting at k. In fact, the only nodes contributing information about the probability of non-adoption are the non-adopter nodes with exactly k exposure. As with $p(k)$, exposure is measured at activation for adopting nodes and at the end of observation interval for non-adopters.

$$v(k) = \frac{\text{all nodes that activate with exposure } = k}{\text{all nodes with exposure } = k} \tag{5.2}$$

This version corresponds to a one-shot experiment where the experimenter rolls a die and activates k peers, then records if the subject activates. The $v(k)$ curve can be interpreted as the observed probability of activation for a random node assigned exposure k.

If it was possible to independently manipulate the exposure level for each node in a laboratory setting we could measure each node's true threshold and construct a conditional probability curve. In this hypothetical controlled laboratory experiment, a node observed inactive with k exposure must have threshold

greater than or equal to k, so the $p(k)$ for true threshold T is:

$$p_{\text{true}}(k) = \frac{T = k}{T \geq k} \tag{5.3}$$

This construction parallels the $p(k)$ for observed exposure at activation.

5.2 $p(k)$ curves and the underlying threshold distribution

Sizable fully connected subgraphs are not common in online social networks, but even small groups of well connected nodes can make distinguishing between simple and complex contagions impossible from exposure at activation. The network structure constrains what we can observe. The extent of the problem can be illustrated via simulation. In the next section, I investigate the relationship among node activation thresholds, observed exposure and observed behavior.

5.2.1 Simulation

The simulations start with a network and a distribution of thresholds over the nodes. The results below reflect node thresholds drawn uniformly at random from the interval [3,9]. The seed cluster is a random focal seed node plus up to 20 of the seed neighbors which have their threshold reduced to zero. All nodes draw their next update time from an exponential distribution. Nodes are loaded into priority queue and visited in order of update time.

As each node is processed from queue, its current exposure is compared against its threshold. If the current exposure is greater than or equal to the threshold, the node is becomes activated and the its activation exposure is logged. Each node draws an interval-until-next-visit from an exponential distribution and re-enters the queue. The simulation proceeds until a target number of nodes activates.

5.2.2 Results

Results from illustrative cases show the correspondence between observes threshold curves and true thresholds is weak. The best results are from a college facebook graph. Graphs with more clustering, like a ring lattice distort the curves even more. The plots below show the $p(k)$ curve (E>=K), the $v(k)$ curve (E==K) and the $p_{\text{true}}(k)$ curve we could construct if we each persons activation exposure could be independently observed (Ind. Obs.). In the $p(k)$ curve, many nodes have exposure much greater than their threshold at time of activation - the maximum threshold in the population was 9 but the exposure at activation can be well over the axis limit of 70.

Figures 5.1, 5.2 and 5.3 show the true and observed $p(k)$ at approximately 10%, 50% and 80% of the nodes active for a simulation on a college Facebook network. Notice that $p(k)$ (E>=K) gets less noisy as the the number of nodes activated increases but generally holds its shape. The probability of adoption at any k within the range of true thresholds increases with nodes active while the probability of adoption at $k > \max(\text{threshold})$ stays relatively constant at about .01 - .015.

In contrast, $V(k)$ (E==K) steadily approaches unity for all k greater than or equal to the maximum threshold as the number of nodes activated increases. The shape of v(k) is sensitive to the proportion of activated nodes.

While it may be tempting to relate the max value of the $p_{\text{true}}(k)$ to the max value of the other exposure curves, Figure 4 shows that this pattern does not hold in general and may be sensitive to graph clustering.

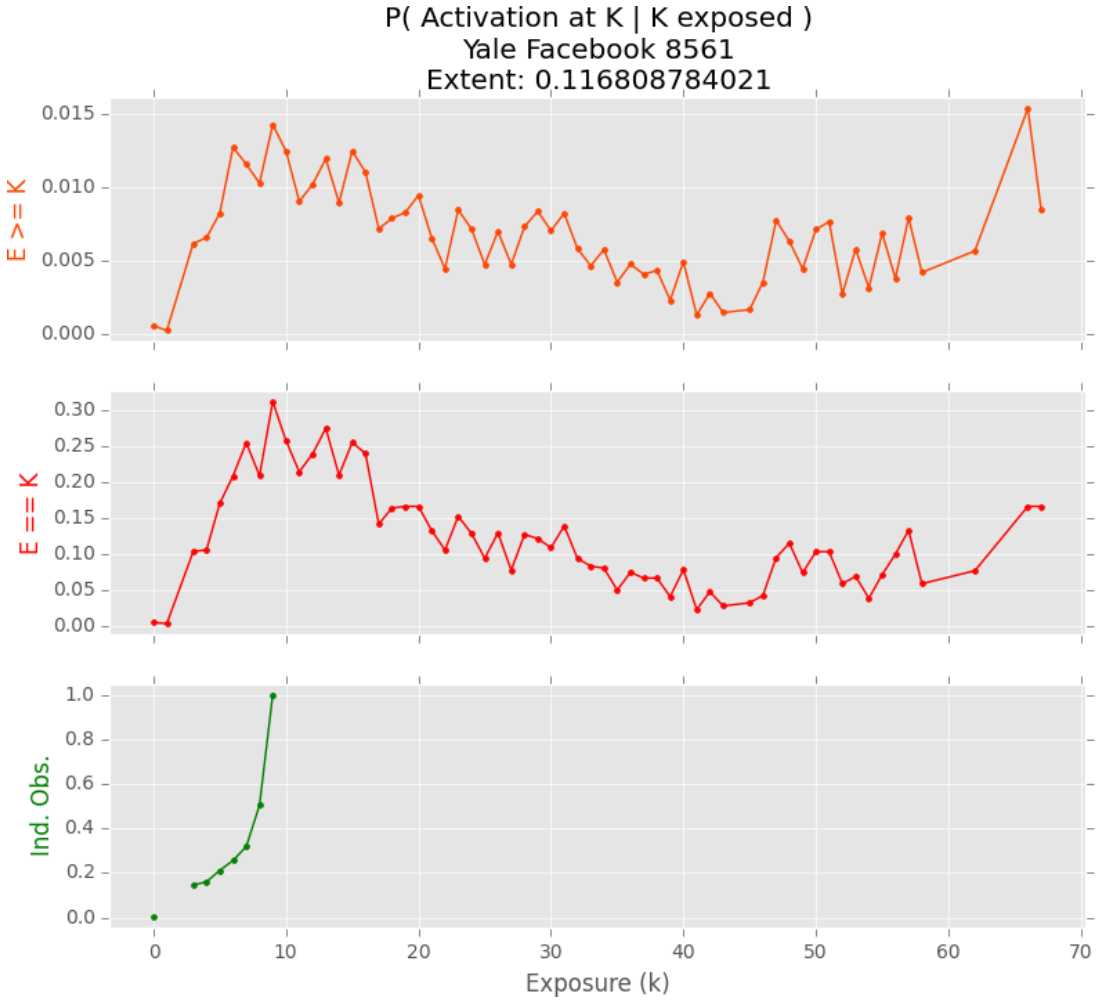


Figure 5.1: Observed $p(k)$ on Yale Facebook network at 10% activation

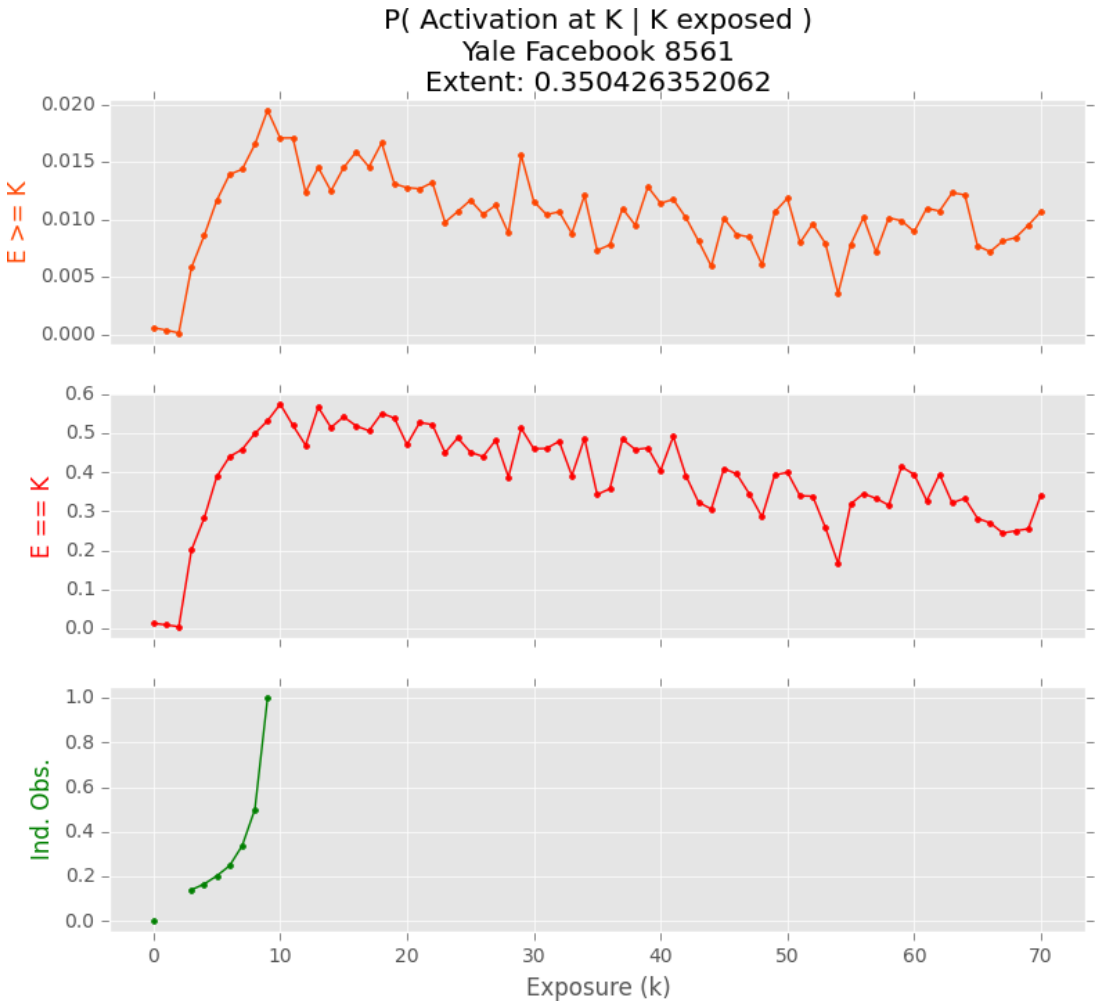


Figure 5.2: Observed $p(k)$ on Yale Facebook network at one-third activation

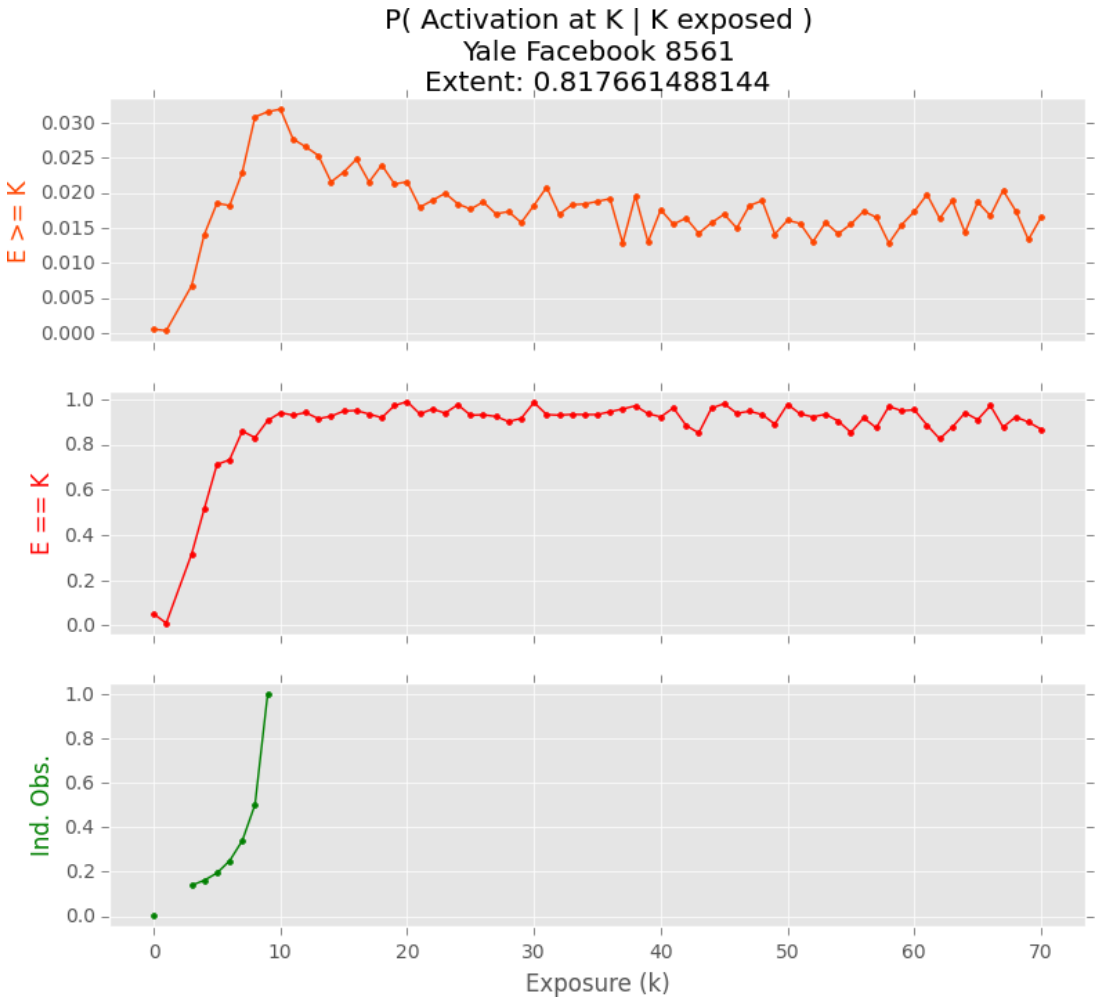


Figure 5.3: Observed $p(k)$ on Yale Facebook network at 80% activation

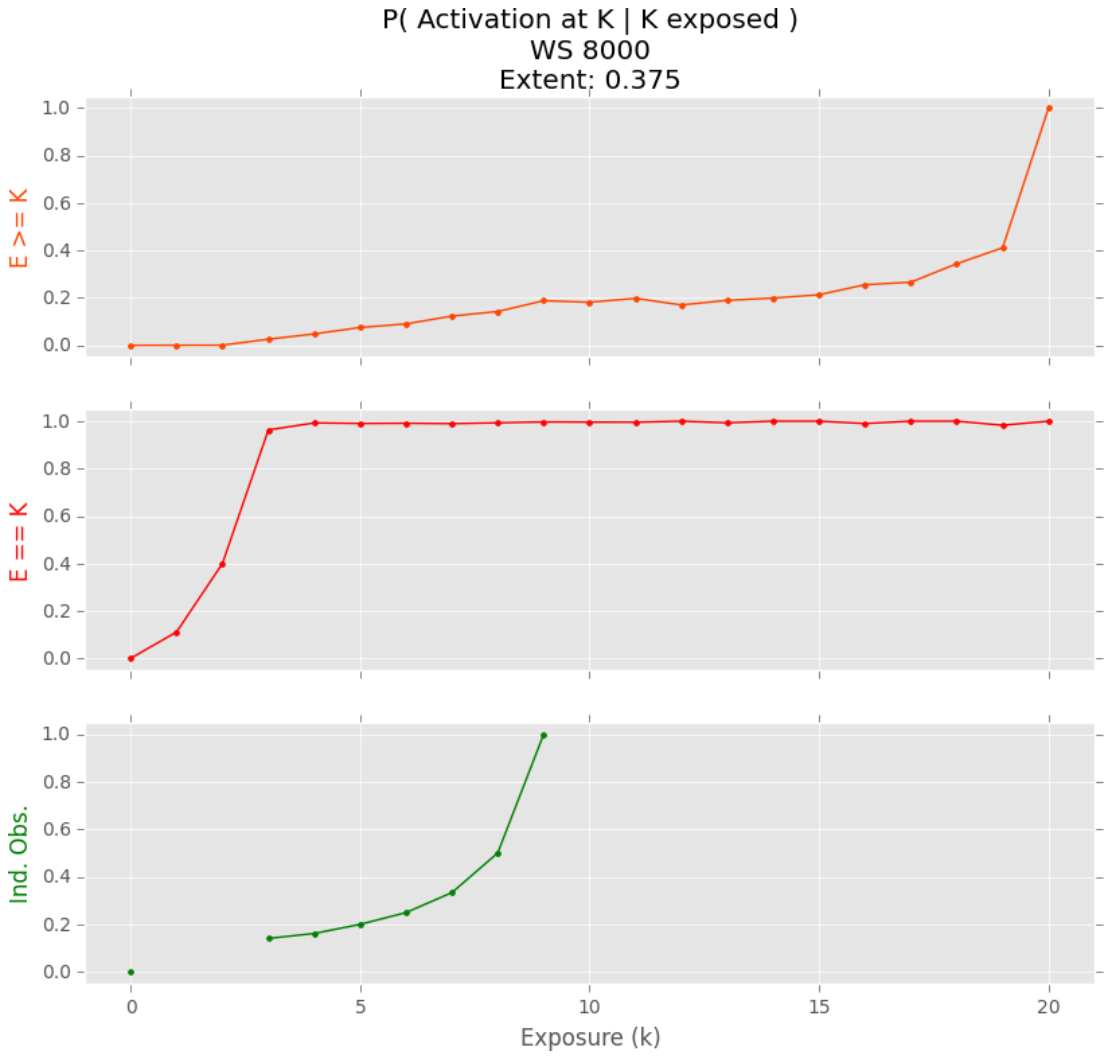


Figure 5.4: Observed $p(k)$ on Watts-Strogatz at one-third activation

5.3 Sources of bias in threshold estimation

Even in this simple simulation, exposure at activation is an imperfect measure of true threshold with potential bias only limited by the node degree. There are three distinct sources of error in observational data; the update interval, the observation interval and the correlated node states.

Because nodes check their status at intervals, their perceived exposure may jump from well below their threshold to well above their threshold between status checks, resulting in an observed exposure at activation that is much larger than the true threshold. The nodes may not have a chance to experience all levels of exposure between zero and their observed exposure at activation.

A related problem comes about when there is some response lag so an individual may experience exposure of $T - 1$ and T but may not activate before the $T + 1$ exposure. With empirical data, we can observe the behavior in response to activation but cannot identify the activating exposure. If exposures occur quickly in comparison to the behavioral response, we observe an activation threshold higher than the true threshold.

The finite observation period creates a counting issue for inactive nodes. Exposure levels for active nodes are computed when the node updates. Exposure levels for inactive nodes are typically computed by looking at their status at the end of the observation period. The update-lag observed in the simulation suggests that at any particular point in time a fraction of inactive nodes will have a current exposure greater than their threshold. These nodes would activate on their next visit but we stop observing before their activation.

Finally, there is an issue of correlated states. Is is impossible to activate nodes asynchronously and also observe the true thresholds for all nodes.

Each of these measurement issues is exacerbated by node degree, because each additional neighbor increases the chance that a node's state will change in the between status updates.

5.4 A partial solution

The magnitude of the bias from update interval and observation period can be reduced if both the nodes highest known exposure while inactive and lowest exposure while active are taken into account when computing the possible thresholds. In data sets where sign-in times are available would provide the highest quality bounds, but other information could be used to model login frequency.

The bias from a finite observation period can be corrected by using the last true observation for each node. Instead of using the node exposure at the end of the observation period, it is more correct to use the node's exposure at time of its last visit. The issue is particularly important for simulation where the spread of the contagion is artificially truncated but could also produce inflated inactive threshold for empirical data with short observation windows.

If thresholds are uncorrelated with degree and update frequency and there is no homophily by threshold, then the change in exposure between status updates is a random variable with respect to true threshold. Some cases will have tight bounds and others will have wide bounds. We can take all cases where the difference in exposure between successive visits is less than some cutoff and this

subset is random with respect to true thresholds, so we can produce an estimate of the pk curve from this sample by computing the percentage of the cases that are active on the second visit at each level of exposure.

Adjusting the count method and selecting cases where the difference between successive visits is less or equal to some cutoff, we can come closer to recovering the curve we would observe with independent observations. High degree with clustering still presents an issue because it is relatively hard to observe small changes in exposure between visits as the number of active nodes in the cluster increases. Adjusting the counting method generally improves the shape of the estimated curves

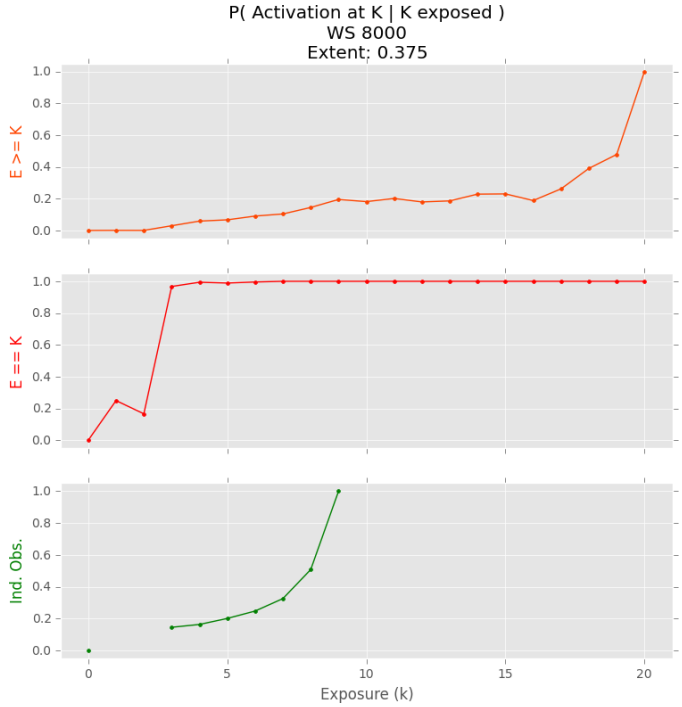
On a ring lattice with mean degree 20 with about one-third of the nodes active, the unadjusted data spans twice the range of the true threshold range, a pattern that is simply the product of the max degree of the network. Figure 5.5b shows that the standard $p(k)$ computation (Cosley et al., 2010) applied to the adjusted and filtered data produces a curve that is a much better approximation of the shape and scale of $p_{\text{true}}(k)$.

On a clustered powerlaw network (Holme & Kim, 2002; Hagberg, Schult, & Swart, 2008), with mean degree 2, the unadjusted data includes k values as large at 120 and the observed $p(k)$ is erratic. For raw data, $v(k)$ appears to be a better match overall. The results for adjusted data, shown in Figure 5.6b, reveal that both methods produce reasonable similar results.

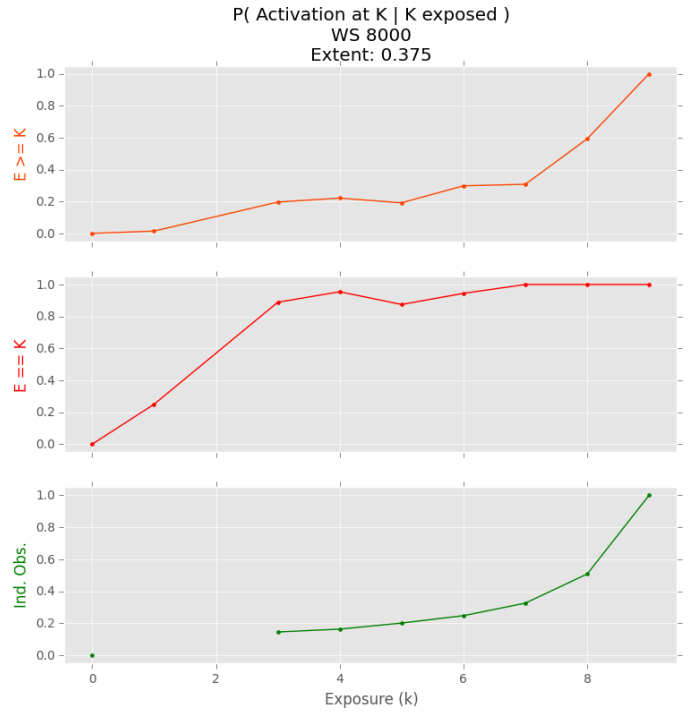
Finally using a college Facebook network, clustering and extreme node degree cause some data sparsity issues and the cases with threshold ranges less than 8 are used to estimate the curves. The adjusted data again produce better

estimates though $v(k)$ and $p(k)$ each deviate in their own way from the expected shape. The curve for $v(k)$ has a more accurate range while $p(k)$ has a more accurate shape.

An adjusted measure might give better results but high node degree and clustering creates a data sparsity issue for the adjusted measure. This does not address the issue with behavioral lag after activation. Contagions that involve substantial unobserved investment before generating an observable behavior require other methods.

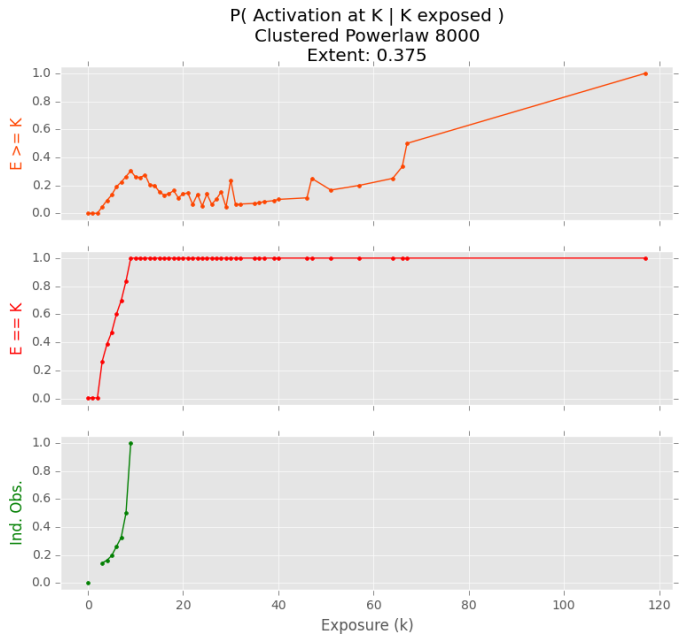


(a) Unadjusted data: note the mismatch in both shape and scale.

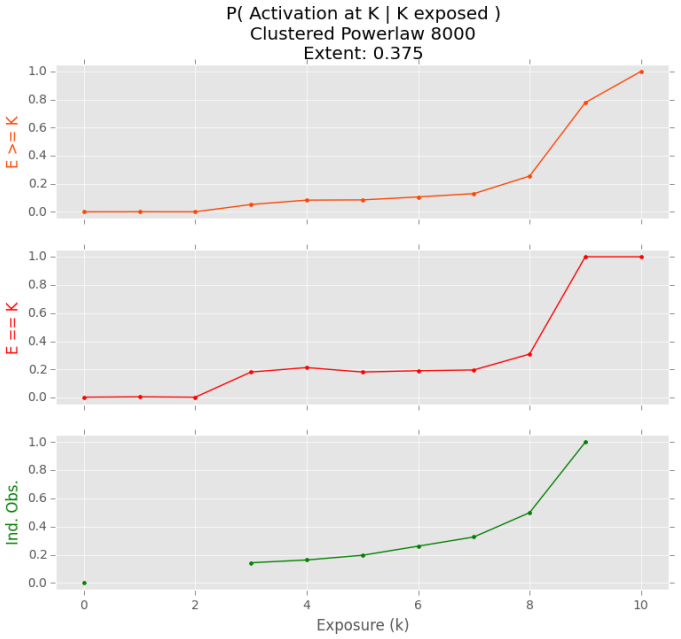


(b) With adjusted counting and cutoff of 1, the top $p(k)$ curve closely approximates the curve for independently measured observations.

Figure 5.5: Attempt to construct $p(k)$ curve for ring lattice with mean degree 20 using unadjusted and adjusted data

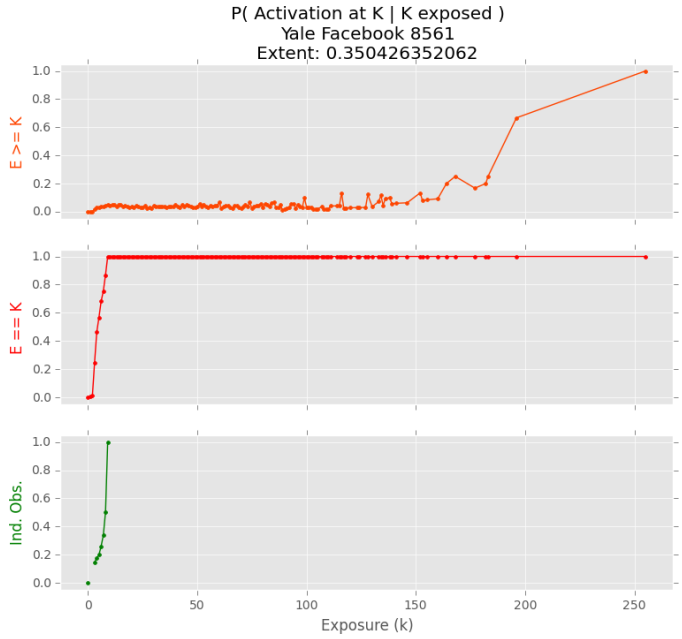


(a) unadjusted

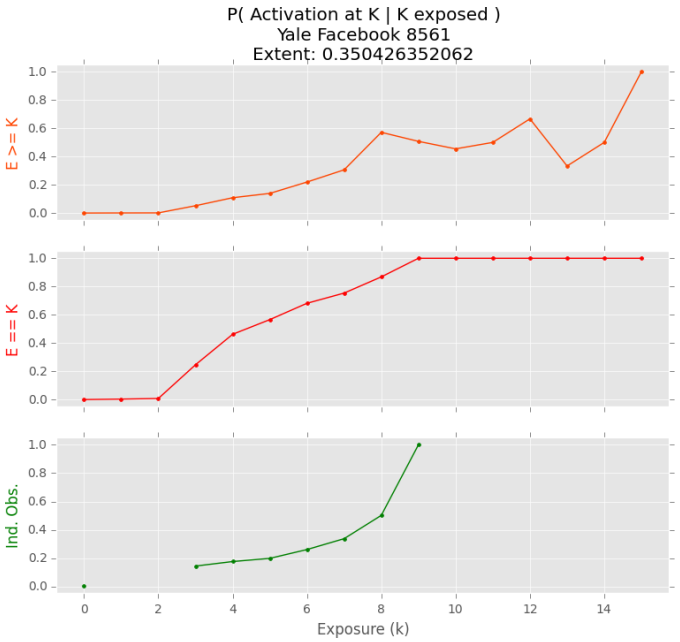


(b) adjusted

Figure 5.6: Attempt to recover $p(k)$ on a clustered powerlaw graph with mean degree 20 using unadjusted and adjusted data. With adjusted counting and cutoff of 2, both $p(k)$ curves approximate the curve for independently measured observations. Data are too sparse for cutoffs less than 2.



(a) unadjusted



(b) adjusted

Figure 5.7: Attempt to recover $p(k)$ on a Yale Facebook graph graph with mean degree 95 using unadjusted and adjusted data. With adjusted counting and cutoff of 8, both $p(k)$ curves somewhat reflect the curve for independently measured observations. Data are too sparse for cutoffs less than 8. The extended range is due to the larger cutoff.

5.5 Discussion

Inferring thresholds from observational data is subject to bias from multiple sources. Changes in the network neighborhood between individual status updates and the correlated nature of node states prevents nodes from experiencing a the full sequence of exposure levels. Using a fixed observation window inflates the number of non-adopoters. Together, these biases make the relationship between observed exposure at adoption and underlying threshold distribution tenuous at best.

Redefining exposure in terms of the exposure levels actually experienced by the individual along with careful tracking of user login times dramatically improves the threshold estimates inferred from exposure levels at activation.

Despite these improvements, there are indications that a hybrid measure might be more accurate in some situations. Exploiting the fact that the true threshold is more likely to be towards the lower end of the range between highest inactive exposure and lowest active exposure could lead to bootstrap or weighted estimates. These corrected estimators could be less sensitive to the larger cutoffs required in networks with high degree and clustering.

CHAPTER 6

CONCLUSION

6.1 Summary

The present work builds on my research effort for Barash et al. (2012) and develops complex contagion theory to better bear on empirical data. The results presented here support a collaborative and ongoing research program in complex contagion with the Social Dynamics Laboratory at Cornell University.

6.1.1 Clusters emergence

I described the dynamics of the complex contagion model identifying the contagion motifs and larger scale dynamics observed during the contagion life-cycle I identified the formation of locally isolated clusters as the mechanism that allows complex contagions to spread faster on networks with some long range ties.

Both the number and the distribution of long range ties is critical to development of such clusters. Even though long range ties do not activate many nodes directly, they are responsible for the establishment of new clusters which expose new groups of nodes. Once the cluster is established contagion proceeds via short range ties in a radial fashion.

The changes in the timing of and frequency of cluster formation correspond to the decrease in time to saturation and the the relationship between cluster emergence and time to saturation holds within network and across networks.

6.1.2 Distribution of tie range

I surveyed empirical social networks and found that the range of long range ties is decidedly shorter than previous models assumed. I implemented an alternative permutation strategy (GEO) that produces more realistic distributions of tie range and established that changes in the distribution of tie range result in qualitative differences in the contagion dynamics.

6.1.3 Clusters and the distribution of tie range

I found that the pattern of spread on GEO permuted lattices is markedly more radial than on MS lattices, even for simple contagions and the role of clusters is significantly reduced. The adoption curves show a pattern of smooth acceleration without identifiable phases for all levels of threshold. The overall effect of GEO permuted long range ties compared to MS permuted ties is to dampen the impact of long range ties. Compared to MS permutation, it takes more long range ties to increase the rate of propagation and to also more to cause a detrimental effect. Even minimally complex contagions are unable to establish viable clusters.

Instead of clusters, contagions in GEO permuted lattices form deep perimeters, a new kind of contagion pattern, where long range ties activate nodes a little beyond the boundary between adopters and non-adopters and then expose the intermediate nodes from both sides.

Though empirical networks may lack a structure that can produce sudden acceleration of the adoption rates endogenously, but they are likely susceptible

to rapid acceleration if clusters are extablished by another mechanism. If spontaneous instigators or current events create small clusters of new adopters, GEO permuted networks are just as capable of the local spread that drives growth once the cluster is established.

6.1.4 Measuring contagion thresholds from observational data

I identify a measurement issue that biases estimates of node thresholds generated from their observed exposure at activation. I show the bias can be reduced by collecting additional data about non-activating exposures.

Inferring thresholds from observational data is subject to bias from multiple sources so the relationship between observed exposure at adoption and underlying threshold distribution tenuous at best. Redefining exposure in terms of the exposure levels actually experienced by the individual along with careful tracking of user login times dramatically improves the threshold estimates inferred from exposure levels at activation.

6.2 Limitations

The complex contagion model considered here necessarily simplifies a complex social process. In comparison with previous models, I consider one additional factor—the distribution of tie range—that influences the dynamics of complex contagion. Studying the impact of this single additional factor in isolation provides significant insight about its consequences for contagion dynamics but may not predict the real world impact of the distribution of long range ties once other

factors are considered.

The model focuses narrowly on one aspect of adoption—exposure from direct network neighbors. Empirical examples of contagion adoption likely involve a mixture of mechanisms, including exposure via mass media broadcasts and more localized individual experiences that reduce thresholds to adoption. A related class of models considers threshold in terms of the proportion of neighbors adopting rather than the number of neighbors. Some types of contagions, particularly those with benefits or costs that scale with fraction of adopters might be better modeled with a different definition of threshold.

A more realistic model would treat threshold as a node level property which is also specific to the contagion. For a given contagion, there would be a distribution of thresholds in the population and different distribution shapes and assortativity among nodes with different thresholds could produce very different outcomes. Another step towards realism would be the introduction of stochastic adoption where threshold influences the probability of adoption as exposure increases. Such stochasticity could increase the probability that long range ties are able to establish new clusters.

The modified procedure for estimating thresholds relies on data about node status updates that is not always available and the requires certain assumptions about the relationship between update frequency and node attributes that might not be plausible in human social networks.

While the present work addresses some important challenges in applying complex contagion theory to empirical data, there are many opportunities for additional research in this area.

6.3 Directions for future work

6.3.1 Examination of complex contagion and dynamics in empirical networks

Much of the motivation for the examination of contagion dynamics in the present work was motivated by the efforts to identify and understand the spread of possible contagions through online social networks. This empirical work relies on user activity and interaction logs to map the spread empirical contagions through social networks. Prior to the present work, the relatively small visible contribution of contagion via long range ties and the tendency for empirical contagions to form only a few locally isolated clusters presented something of a puzzle as it did not match our expectations. With new insight about the impact of the distribution of long range ties the next step it to validate and extend the results in empirical networks.

These efforts include:

1. Simulations of complex contagion on empirical networks to compare the dynamics with the observed patterns of spread. Although specific networks only occupy only a single point in the parameter space, the emergence of clusters should be rare for contagions of all thresholds.
2. Developing a measure of perimeter depth, as this is an important mechanism that facilitates faster spread in networks with more realistic tie range.
3. Examine why some popular contagions (hashtags) do develop clusters while many popular tags do not develop sizable secondary clusters.

6.3.2 Bootstrap approach to detect higher threshold contagion

Some nodes will adopt with high exposure levels because they have high thresholds and others will adopt with high exposure because they are constrained by network structure and happened to update later than their neighbors. The high exposure attributable to network structure will be present even if nodes update in a random order. Comparing the distribution of exposures at activation under the random activation assumption with the observed distribution of exposures will highlight notable deviations in exposure at adoption that are not readily explained by network structure. A contagion should have fewer activations at zero exposure and a complex contagion should have more adoptions with exposure greater than one. Even if thresholds for individual nodes cannot be measured in an unbiased way, the bootstrap should pick up patterns related to the overall distribution of thresholds among the first X adopters.

6.3.3 Model threshold as a function of node attributes

The suggested improvements in threshold measurements from observational data depend on node threshold being random with respect to the factors that contribute to a node's probability of being observed with a small difference between nonactive and active exposure. These factors include network properties like degree as well as personal attributes which may plausibly impact the frequency of site visits relative to peers. In order to combat this dependency, the threshold can be modeled as a function of note attributes and estimated from the data to produce weighted threshold estimates that account for differential probability of observing the different groups. This paper, co-authored with George

Berry is targeted towards Sunbelt 2016

6.3.4 Early detection of accelerating adoption

Cluster emergence may still be an import part of real world contagion processes even if they are not created endogenously. This line of research would investigate the relationship between adoption rate accelerations and cluster emergence in real data. If clusters are found to be important for some contagions, then careful examination of the non-structural mechanisms that produce new clusters could provide a method for the early detection of virality.

6.3.5 Expanded examination of empirical tie range

The present survey focused on the empirical distribution tie range to answer a specific theoretically motivated question about complex contagion dynamics. I have examined the range of ties in empirical and simulated networks which is the same as the examining the second shortest path between nodes that share an edge. While the distribution of long range ties is demonstrated to be related to propagation dynamics, further work could examine redundancy of ties of various lengths. The problem of computing the k-shortest paths between nodes has inspired the creation of several algorithms (Eppstein, 1994; Papaefthymiou, 1997; Li et al., 2006; Kao et al., 2011; Zhang & Nagamochi, 2012; Wu, 2013). The number of paths of each length could be related to the concept of local bridge *width*, introduced by Centola and Macy (2007), and provide further insight about a network's potential to spawn new areas of contagion.

6.3.6 Problematic implications of edge swap permutation

The double edge swap has the appealing property of changing the graph structure while holding the number of edges and the degree of each node constant. It also has a side effect of producing two long range ties whose endpoints are in similar neighborhoods. While these near-parallel edges are not enough to create a bridge on their own, they do create a structure that might create qualitative differences in the contagion dynamics that do not appear in empirical networks. Future work could examine the distribution of long range ties over nodes to measure the extent to which nodes with long range ties tend to be connected to or near each other.

6.3.7 Hybrid or weighted threshold estimates

Careful accounting of user exposure levels over time produce upper and lower bounds on a user's threshold. The upper bound is limited by node degree which can vastly overestimate the threshold. Bootstrap or weighted estimates could exploit the fact that the true threshold is more likely to be towards the lower end of the range. These corrected estimators could be less sensitive to the larger cutoffs required in networks with high degree and clustering.

6.3.8 Model attention to improve threshold estimates

Information about non-activating exposures can dramatically improve the estimates of node threshold but many sources, like Twitter, lack information about

user sign-on activity. The user activity log only contains information about content submitted the the sign. Nevertheless, post frequency may be sufficient to narrow down exposure. Beyond confirmed activity, the diurnal and weekly patterns of posting and responses to others posts may provide enough information to produce a model of times a user is likely to be online. With a sufficient model of attention to inform estimates of exposure the user threshold can be estimated at smaller granularity.

BIBLIOGRAPHY

Barash, V., Cameron, C., & Macy, M. (2012, March). Critical phenomena in complex contagions. *Social Networks*, *34*(4), 451–461.

Centola, D. (2010, September). The spread of behavior in an online social network experiment. *Science (New York, N.Y.) 329*(5996), 1194–7.

Centola, D., Eguíluz, V. M., & Macy, M. W. (2007, January). Cascade dynamics of complex propagation. *Physica A: Statistical Mechanics and its Applications*, *374*(1), 449–456.

Centola, D. & Macy, M. W. (2007). Complex Contagions and the Weakness of Long Ties. *American Journal of Sociology*, *113*(3), 702–734.

Coleman, J., Katz, E., & Menzel, H. (1966). *Medical innovation: a diffusion study*. Indianopolis: Bobbs-Merrill Co.

Cosley, D., Huttenlocher, D., Kleinberg, J., Lan, X., & Suri, S. (2010). Sequential influence models in social networks. *ICWSM101530*, 26–33.

Csardi, G. & Nepusz, T. (2006). The igraph software package for complex network research. *InterJournal*, *Complex Sy*, 1695.

Eppstein, D. (1994). Finding the k shortest paths. *Proceedings 35th Annual Symposium on Foundations of Computer Science*.

Ghasemiesfeh, G., Ebrahimi, R., & Gao, J. (2013). Complex Contagion and the Weakness of Long Ties in Social Networks: Revisited. *Proceedings of the Fourteenth ACM Conference on Electronic Commerce*, *1*(212), 507–524.

Grannovetter, M. (1973). The strength of weak ties. *The American Journal of Sociology*, *78*(6), 1360–1380.

Granovetter, M. (1978). Threshold models of collective behavior. *American journal of sociology*, *83*(6), 1420–1443.

Granovetter, M. (1983). The strength of weak ties: A network theory revisited. *Sociological theory*, *1*(1983), 201–233.

Hagberg, A. A., Schult, D. A., & Swart, P. J. (2008). Exploring network structure, dynamics, and function using NetworkX. *Proceedings of the 7th Python in Science Conference (SciPy 2008)*, (SciPy), 11–15.

Holland, P. W. & Leinhardt, S. (1978, November). An Omnibus Test for Social Structure Using Triads. *Sociological Methods & Research*, *7*(2), 227–256.

Holme, P. & Kim, B. J. (2002, February). Growing scale-free networks with tunable clustering. *Physical review. E, Statistical, nonlinear, and soft matter physics*, *65*(2 Pt 2), 26107. doi:10.1103/PhysRevE.65.026107

Kao, K. H., Chang, J. M., Wang, Y. L., & Juan, J. S. T. (2011). A quadratic algorithm for finding next-to-shortest paths in graphs. *Algorithmica (New York)*, *61*(2), 402–418.

Kleinberg, J. (2000). The Small-World Phenomenon: An Algorithmic Perspective. *Proc. ACM Symposium on Theory of Computing*, 1–14.

Leskovec, J., Adamic, L. A., & Huberman, B. A. (2007, May). The dynamics of viral marketing. *ACM Transactions on the Web*, *1*(1), 5–es.

Li, S., Sun, G., & Chen, G. (2006). Improved algorithm for finding next-to-shortest paths. *Information Processing Letters*, *99*(5), 192–194.

Maslov, S. & Sneppen, K. (2002). Specificity and stability in topology of protein networks. *Science (New York, N.Y.) 296*(5569), 910–913.

Meyer, D., Zeileis, A., & Hornik, K. (2015). *vcd: Visualizing Categorical Data*. R package version 1.4-1.

Milgram, S. (1967). The Small-World Problem. *Psychology Today*, *1*(1), 61–67.

Mislove, A., Marcon, M., & Gummadi, K. (2007). Measurement and analysis of online social networks. *Internet measurement*.

Newman, M. E. J. (2003). The structure and function of complex networks. *Dialogues in clinical neuroscience*, *45*(2), 167–256.

Papaefthymiou, M. C. (1997). Information Processing. *Information Processing Letters*, *63*(97).

Park, P. (2016). *The Strength of Long Ties* (Doctoral dissertation). Cornell University.

Romero, D., Meeder, B., & Kleinberg, J. (2011). Differences in the mechanics of information diffusion across topics: Idioms, political hashtags, and complex contagion on Twitter. In *Proceedings of the 20th www* (pp. 695–704).

Taylor, D., Klimm, F., Harrington, H. a., Kramár, M., Mischaikow, K., Porter, M. a., & Mucha, P. J. (2015). Topological data analysis of contagion maps for examining spreading processes on networks. *Nature Communications*, *6*, 7723.

Traud, A. L., Mucha, P. J., & Porter, M. a. (2012). Social structure of Facebook networks. *Physica A: Statistical Mechanics and its Applications*, *391*(16), 4165–4180.

Travers, J. & Milgram, S. (1969). An Experimental Study of the Small World Problem. *Sociometry*, *32*(4), 425–443.

Valente, T. (1995). *Network models of the diffusion of innovations*. Cresskill, NJ: Hampton Press.

Watts, D. J. & Strogatz, S. H. (1998, June). Collective dynamics of 'small-world' networks. *Nature*, *393*(6684), 440–2.

Weng, L., Menczer, F., & Ahn, Y.-Y. (2013). Virality prediction and community structure in social networks. *Scientific reports*, *3*, 2522.

Wu, B. Y. (2013). A simpler and more efficient algorithm for the next-to-shortest path problem. *Algorithmica*, *65*(2), 467–479.

Zhang, C. & Nagamochi, H. (2012). The Next-to-Shortest Path in Undirected Graphs with Nonnegative Weights. *Proceedings of the Eighteenth Computing: The Australasian Theory Symposium*, *128*, 13–20.