INFORMATION CONTAGION IN TEMPORAL HUMAN NETWORKS WITH HETEROGENEOUS SUSCEPTIBILITY

Milo Zappa Trujillo

Submitted in Partial Fullfillment of the Requirements
for the Degree of

MASTER OF SCIENCE

Approved by:
Dr. Jianxi Gao, Chair
Dr. Sibel Adalı, Co-Chair
Dr. Buster Holzbauer

Department of Computer Science
Rensselaer Polytechnic Institute
Troy, New York

[May 2020]
Submitted March 2020

CONTENTS

LIST OF TABLES

LIST OF FIGURES

ACKNOWLEDGMENT

I extend my sincerest gratitude to my thesis advisor, Dr. Jianxi Gao, for his expertise, guidance, and patience. He introduced me to a new field, accepted me as a thoroughly inexperienced researcher, and gave direction towards both my first publication and thesis.

Dr. Sibel Adalı, I cannot thank you enough. Since I first joined your lab four years ago you have been my mentor, lifeline, and role-model in so many aspects of life. I am excited for what the future holds, but will always miss my time at RPI with you.

I also thank Dr. Buster Holzbauer for his role on my defense committee, his encouragement and constructive feedback with this thesis, and his tutelage over the past two years.

I additionally thank Dr. Xueming Liu and their student Ke Wu for their in-depth feedback across many earlier iterations of this work.

Finally, thank you to all the other members of the RPI Trust lab, especially Ben Horne, Maurício Gruppi, and Clare Arrington, who have been unwaveringly supportive, and without whom I would feel wholly out of place in academia.

ABSTRACT

Information Cascades are propagation patterns in human networks where information such as political propaganda, rumors, or fashion trends, emanate from one or more starting points and are either accepted or rejected by each connected member of a community. The decisions made by individuals either govern a large ripple through the community, or prevent significant changes from occurring, making Information Cascades of avid interest within economics and network science. Since Information Cascades have similar behavior to the well studied field of epidemiology, many Cascades are modeled with disease propagation models like the Susceptibility, Infection, and Recovery (SIR) model. In this thesis we argue that such generic models are insufficient for predicting information spread in human behavioral networks, because they do not represent temporal availability of communication and individual variations in susceptibility to peer-pressure. We propose a new model wherein agents have heterogeneous activity periods and activation thresholds, representing individual susceptibility to peer-pressure. By applying this model to networks with a range of topologies, as well as real-world networks, we are able to more accurately represent human behavior in a variety of communities.

CHAPTER 1
INTRODUCTION

1.1 Background

Global Cascades are substantial propagating changes throughout a network triggered by small initial perturbations [1]. These cascades can describe a wide variety of physical and social phenomena, from rippling failures in an electrical grid [2], to waves of stock-market sales [3], and the spread of computer viruses [4]. Studying cascades has the potential to predict and influence behavior in a number of systems, from limiting the spread of infectious diseases to promoting the dissemination of information. In this thesis we model information spread in human social networks, with particular consideration for individual susceptibility to peer pressure and individual activity patterns. We examine cascades starting from both a single point, as in most previous studies, and cascades starting from several random or chosen locations, better representing behavior in information campaigns such as advertisement and propaganda. We study cascade behavior across a variety of network topologies, endeavoring to better understand the role of community structure in information spread, and the interplay between community structure and individual behavior.

In a social interaction network, agents communicate with their neighbors to pass information. In most previous studies, the network topology is often assumed to be fixed [5], and the behavior patterns of each agent are assumed to be homogeneous. In these models, agents have a fixed susceptibility to new information based on pressure from their neighbors, and over time may be "infected" with information (terminology inherited from academic literature on epidemic modeling, such as the Susceptibility, Infection, and Recovery model [6]), and begin propagating information to their neighbors [7]. Assumptions about homogeneous agent behavior and fixed social relationships are unrealistic in real-life scenarios leading to a need for models that are more heterogeneous. In some engineered systems, this heterogeneity can be directly measured, such as in the fault-tolerances of relays in an electric-grid [2]. Human behavior in networks is more challenging to quantify and measure, but it is widely accepted that people have varying susceptibility to peer pressure and differing levels of social activity.

More recent studies account for heterogeneous behavior in human networks, focusing

on network topology [8], individual susceptibility to information [9], and time-based activity periods [10]. For instance, Crucitti, Latora, Marchiori, *et al.* [8] tested network robustness to intentional and accidental damage on random and scale-free (SF) networks. Li, Wang, Sun, *et al.* [11] represented individual heterogeneity as different probabilities of leaving the network in the event of cascading crashes, and show that changes in this individual behavior greatly change the system behavior in network collapses. Watts [9] provided a simple quantitative representation for heterogeneous susceptibility to influence: each agent is given a susceptibility level we call an *infection threshold*[1] from a normal distribution, representing individual personalities. Notably, this model assumes information susceptibility is based solely on the number of affected peers and does not consider some peers to be more influential than others, nor does it consider the possibility of agent susceptibility changing over time.

Other studies consider the effect of time on network structure. Holme and Saramäki [12] introduced a concept of *temporal networks*, where edges are active only at specific periods. Belykh, Belykh, and Hasler [13] created temporal networks that add and remove edges while maintaining small-world network characteristics, which could be used to model social communities restructuring over time. By contrast, Akbarpour and Jackson [10] proposed a subclass of temporal networks with a fixed topology describing social behavior. In this model, edges representing social interactions are constant, but may only be enabled during particular *activity periods*, representing activity patterns such as a workday or school schedule.

Gomez-Rodriguez, Leskovec, and Schölkopf [14] develop an alternative contagion-inspired methodology where the network topology is not known before simulation. Agent infection is observable as the information spreads, but the route of spread (edge between agents) is unknown, meaning individuals are unaware of where they first learned information from. The researchers attempt to predict the edges of the network based on infection, and infer which links impede or facilitate information spread. This is similar to real-world studies on communities with publicly identified members, but obscured social ties and hierarchies [15].

Sobkowicz [16] creates an artificial social network model where leaders have disproportionate influence on the beliefs of their followers. The authors later extend this study to create an opinion dynamics simulation based on real-world data from Polish politics [17].

[1] In previous related papers this is called an "activation threshold". We rename the term to avoid confusion with the similar terminology "activity period"

Galam [18] creates an alternative model representing opinion dynamics in two-party competitive politics. In their system, small groups meet to discuss politics, and change their political alignment based on the majority opinion in the group. Agents have heterogeneous collaborative behavior, and can either conform to the majority opinion, remain inflexible regardless of majority, or reject the group opinion as contrarians. By creating many of these group interactions over time, a political party's dominance spreads through the larger community, ending in their opinion spreading throughout the majority of the network.

Other researchers such as Borge-Holthoefer, Perra, Gonçalves, *et al.* [19] begin with real-world time lapse data rather than simulations. This study applies symbolic transfer entropy to measure unidirectional and bidirectional flow of information about a collective action on Twitter, retroactively showing how cascades start from a small number of committed "driver" users and grow into broader decentralized and symmetric information networks.

1.2 Contributions of This Thesis

In this thesis, we create an agent-based cascade model with different susceptibility of agents on temporal networks, utilizing a range of network topologies from random to SF, including simulated networks and sampled real-world communities. Taking temporal activity patterns and heterogeneous susceptibility into consideration brings models a step closer to the real process of information spreading in human interaction networks than prior representations, which assume fixed topologies and homogeneous susceptibility or consider only one variability in human behavior at a time.

We simulate cascades beginning from both a single starting point, and a minimum threshold multi-start scenario. In the former, a cascade begins at a single agent and we determine the conditions under which the cascade will spread based on the susceptibility and activity of neighbors. In the multi-start case, the network topology and neighbor susceptibility is fixed, and we determine how many agents must begin "infected" to guarantee the spread of the cascade through the majority of the network. In the multi-start scenario we represent cascades with both random and explicitly-targeted high-degree starting locations. This flexibility allows us to represent the organic spread of social information and external influence campaigns such as we might see in propaganda or advertising.

We find that more scale-free networks have fewer cascades that spread throughout the network, but more cascades that spread beyond initial starting regions. Heterogeneity

among activity times deters cascade spread, while heterogeneity among susceptibility produces weak-points that facilitate cascade spread. We further discover that scale-free networks are vulnerable to targeted attacks on hubs, but that this vulnerability exists in networks with a highly power-law degree distribution and extremely high-degree hubs. Communities become dramatically less vulnerable with even slight random perturbations to network linking, and in most scenarios random attacks are only slightly less effective than targeted ones. Simulations on real-world network samples show that our results generalize beyond artificially constructed topologies, but real-world networks may contain loosely linked sub-communities that further isolate cascades, emphasizing the importance of network bridges.

CHAPTER 2
PROPAGATION MODEL

We have created an agent-based model with n agents, bidirectionally linked together either randomly (according to the Erdős-Rényi (ER) model [20] with an arbitrary linking probability), or in a Scale-Free (SF) configuration according to a simple power-law degree distribution with an arbitrary exponent. In both cases, we examine only the giant connected component, ignoring unreachable agents. The model proceeds in discrete time steps for simplicity. At each timestep, any agents "contagious" with information cascade will attempt to spread the cascade to each of their neighbors. This spread is limited by the infection threshold of the target agent and whether they are currently active. Each infected agent spreads to all neighbors synchronously, such that any agents "infected" in timestep t do not change their state until the start of time $t + 1$.

For each agent, we assign an *infection threshold*, which varies from 0 to 1, representing the percentage of neighbors that must be infected before a cascade that can spread to the agent. For example, an infection threshold of 0 indicates that *any* infected neighbor will infect the agent. In contrast, an infection threshold of 1 means that *all* neighbors must be infected before the agent will succumb to a cascade. Depending on the community behavior being simulated, the infection thresholds are either assigned homogeneously throughout the network or assigned according to a uniform distribution $[0, \theta]$.

Once an agent is infected, it is permanently infected, without any concept of "recovery" as in the SIR model [6]. It is because recovery has a precise meaning in a disease model, but has no obvious parallel in an information-spreading scenario. However, agents are only considered "contagious" (spreading the cascade to peers) for T time steps, representing the time during which the cascade-spread information is novel. This contagious period can be calibrated to reflect any real-world cascade being represented but must be set to a finite number, as the simulation concludes when no agents remain contagious. In our trials, we set a contagious period of 10, determined experimentally to be a suitable baseline. Increasing the contagious period beyond 10 did not significantly further the spread of cascades, so at this value, the cascade is limited primarily by agent susceptibility and activity, as we desired, rather than being limited by an arbitrarily-set constant.

Each agent has a activity state: "active" or "inactive". Information can only spread between two active agents. The transition between these two states is modeled as a markov chain such that if an agent is active at timestamp T, it will active at time $T + 1$ with probability $1 - p_i$ and inactive with probability p_i [10]. An agent inactive at time T will become active with probability q_i and will remain inactive with probability $1 - q_i$. This relationship is visualized in Figure 2.1, below. The probability of active or inactive follows the following relationship:

$$\lambda p_i = (1 - \lambda) q_i \tag{2.1}$$

where the parameter λ describes the total ratio of time that the agent is active. Increasing the activity period increases the susceptibility of an agent, but we find in section 3.3 that the effect is significant only when the network topology and infection thresholds are near a critical point and would otherwise allow a cascade to spread.

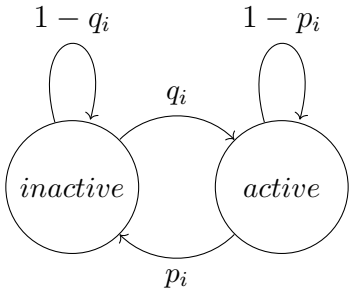


Figure 2.1: Activity Markov chain of agents, adapted from Akbarpour and Jackson [10]. Represents the probabilities of each agent transitioning between active and inactive states.

Furthermore, the transitions between activity states may follow different patterns. In this work, we study four activity patterns as described in Ref. [10]: (1) Poisson agents that randomly switch states every timestep with a specified probability distribution for being active at any step; (2) Sticky agents that remain active or inactive for long stretches to maintain their specified activity percentage, representing sequences like a 9-5 work day; (3) Reversing agents that alternate between active and inactive as frequently as possible while maintaining their specified activity percentage, representing periodic behavior like checking a phone for new messages every few minutes; (4) AlwaysOn agents active at all times, serving as a control group to remove activity patterns as an influence on cascade success.

CHAPTER 3
SIMULATION METHODOLOGY

3.1 Infection Threshold Parameter Sweeps

We initially examined the effect of increasing infection thresholds on homogeneous networks of agents with a specific activity type. For each simulation, we created random (Erdős-Rényi [20]) and Scale-Free networks [21] with the parameters shown in Tables 3.1 and 3.2. We generated each network with a single agent type (AlwaysOn, Poisson, Sticky, or Reversing), and initially a homogeneous infection threshold. In each simulation, a single randomly-selected node is infected at a time step 0, and all other agents are susceptible. The simulation ends when no agents remain contagious. For each configuration, we generate and simulate 300 networks, so anomalous network topologies or cascade starting points are unlikely to affect the median result.

Table 3.1: Network parameters for initial simulation on Erdős-Rényi communities.

Nodes	Avg. Degree	Avg. Activity
1000	5	60%

Table 3.2: Network parameters for initial simulation on scale-free communities.

Nodes	Edges Per Node	Clustering Exponent	Avg. Activity
1000	2	2	60%

We then simulated networks with heterogeneous infection thresholds. For Figure 4.2 the x-axis represents the *upper bound* infection threshold, meaning each agent has a uniformly random infection threshold between 0.0 and the value of the x-axis.

We graph each simulation (Figures 4.1 and 4.2) as a data point, with color denoting the density of results. This allows us to represent both patterns of simulation behavior and exceptions to those patterns. An infection threshold of 0.001 implies the cascade never spread past the initial starting node, while a result of 1.0 means the cascade has reached the entire network. Results are on a color gradient such that yellow data points indicate many results in that region, while indigo indicates few data points in the region. This visualization

is ideal for describing bimodal distributions where some cascades spread through 100% of the network and some through 0% of the network, because a simple mean could misleadingly report than an average cascade spread through 50% of the network.

3.2 Network Topology Sweep: Confirmation Study

We confirm our results from Section 3.1 across a range of ER linking probabilities and SF clustering coefficients to show that the observed cascade behavior is not anomalous. We test networks of 1000 agents with Poisson activity patterns, varying both the network parameters and infection thresholds in Figure 4.3. This sweep also provides an analog for comparing urban and rural environments with different densities of social connections and different levels of heterogeneity.

3.3 Activity Period Parameter Sweep

We tested a variety of activity times for each agent behavior. As in Section 3.1 we tested on random and scale-free networks, with homogeneous and heterogeneous infection thresholds. For homogeneous agents we set a fixed infection threshold of 0.15, and for heterogeneous agents we set an upper-bound threshold of 0.5. These thresholds correspond with the phase transitions in Figures 4.1 and 4.2, allowing us to examine the parameter range where activity periods are likely to have the most significant effect.

We performed a parameter sweep from 0% agent activity (never active, cannot be infected) to 100% agent activity (active at every time step, infection limited only by infection threshold). In these simulations (Figures 4.4 and 4.5) the x-axis represents the percent of time agents are active, and the y-axis again represents the percent of the network a cascade spread through. Note that agent activity is not synchronized; that is, two agents with 80% activity are both active at 80% of all time steps, but not necessarily the *same* timesteps.

3.4 Minimum Starting Infections: Random and Targeted Attack

Lastly, we examine a scenario where cascades begin from a variable number of agents and network parameters are fixed. We simulate both a case where cascades start from multiple random agents, and one where cascades start from a number of the highest-degree agents, and determine how many agents must begin infected for the cascade to spread through

at least 95% of the network. This methodology has been used to simulate random equipment failures and malicious attacks in distributed load networks such as power grids [22], but is equally appropriate in an information-spreading context.

In these trials we simulate 100 networks of 1000 Poisson agents with 2000 total edges for each set of network parameters, and increase the number of starting agents until the cascade dominates the network. These iterations limit the effect anomalous network topologies or unusual random starting locations can have on results.

3.5 Real-World Networks

To confirm our results outside of artificially-constructed networks, we apply our simulation to two real-world networks: An academic collaboration network within General Relativity and Quantum Cosmology [23], and a sampling of Facebook friend circles [24], both publicly available through the Stanford SNAP repository [25]. In both networks we consider only the giant weakly-connected component, ignoring isolated sub-communities a cascade would be unable to reach.

Table 3.3: Network attributes for real-world SNAP networks. Clustering exponent determined by the Kolsmogorov-Smirnov test [26].

Network	Nodes	Edges	Avg. Degree	Clustering Exponent
GrQc Citation	4158	13422	6.456	2.063
Facebook Friends	4093	88234	43.691	2.510

The citation network represents authors in the General Relativity and Quantum Cosmology category of ArXiV. Each edge represents one or more co-authored papers. The Facebook friend-circle network consists of nodes representing individual users, and edges representing a bidirectional friend status. Neither network includes duplicate or self-edges.

Network attributes for both networks are shown in Table 3.3. The citation network has an average degree similar to our simulated networks, and is extremely scale-free. The Facebook friend network has a dramatically higher average degree than what we have tested before, and has a degree distribution further from a perfect power law, but is still well within scale-free bounds.

We conduct a parameter sweep from an infection threshold of 0.0 to 1.0 with increments of 0.01, for a total of 101 testing conditions. For each condition, we test 300 random assignments of infection thresholds, activity periods, and starting infected agent, according to the

methodology in Section 3.1. This provides thorough test coverage for two fixed, real-world network topologies.

For reader convenience, we provide visual renderings of both the real-world networks Figures 3.1 and 3.2, as generated by a stochastic force-atlas clustering in Cytoscape [27].

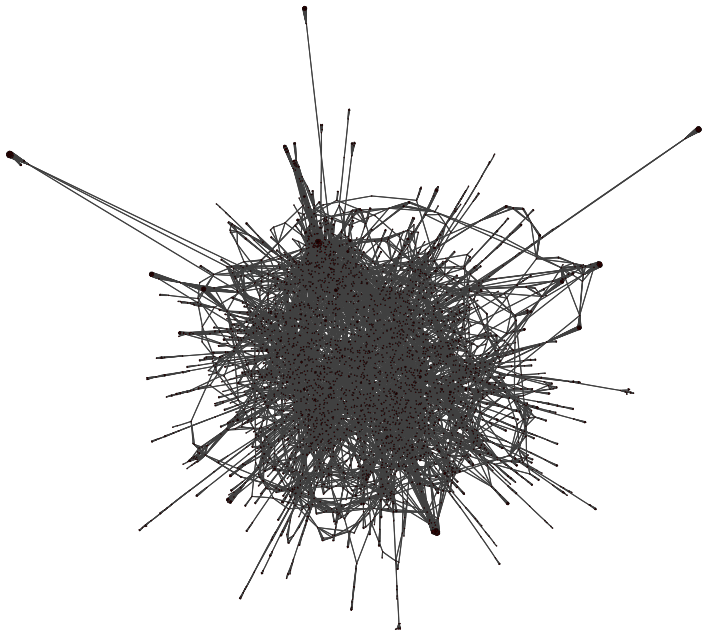


Figure 3.1: Cytoscape render of collaboration network [23] using an edge-repulsive weak clustering layout. Nodes represent authors, edges represent co-authorship of a paper. Most authors are part of the central cluster with stereotypical SF attributes.

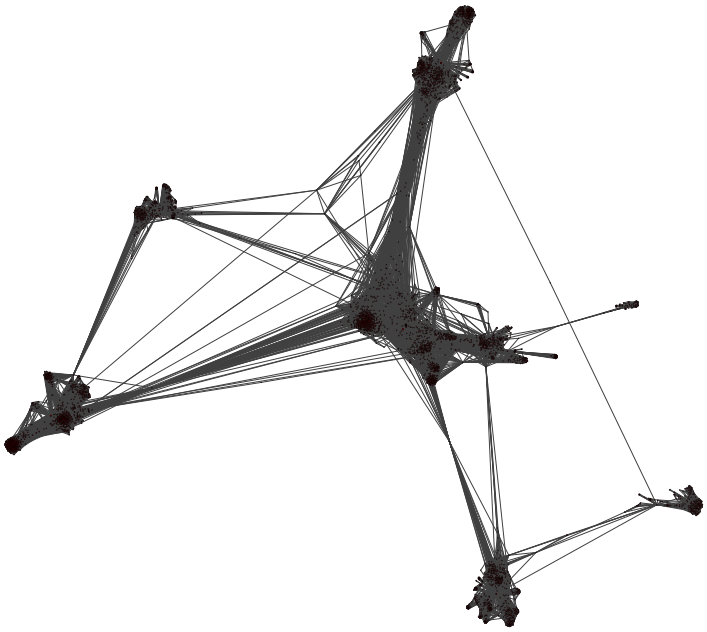


Figure 3.2: Cytoscape render of Facebook friend network [24] using an edge-repulsive weak clustering layout. Nodes represent Facebook users while edges indicate a bidirectional friendship status. SF degree distribution is present, but nodes are clearly distributed across several distinct friend-groups, with few bridges in-between.

CHAPTER 4
RESULTS

4.1 Information Spread in Social Network with Homogeneous Infection Thresholds

We first created networks with homogeneous infection thresholds, so that every agent in the network had the exact same infection conditions. In these conditions, most information cascades spread throughout the network when the infection threshold is lower than a critical point, or restrict the information to close to the initial starting point if the infection threshold is above this critical value, showing a discontinuous first order phase transition on the final infected fraction of agents as the infection threshold increases. Both the activity patterns and network topology have a small effect on the success of an information cascade when agents have homogeneous traits.

In all homogeneous activation networks, cascades generally succeed (spread through the at least 95% of the network) when the infection threshold is below the average degree of agents in the network, and fail shortly above this same point (Fig. 4.1). Given a constant infection threshold, the only variation in cascade susceptibility comes from the degree of each agent. The lower degree an agent has, the higher susceptibility that agent has. For instance, an agent with a degree of three and an infection threshold of 0.3 requires only one infected neighbor to activate, while an equivalent agent with a degree of nine requires three infected peers to activate. Without early susceptible low-degree neighbors, the cascade cannot build a sufficient base to capture higher-degree peers. This effect is particularly pronounced in random networks, as shown in Figures 4.1a to 4.1d, where all agents have close to a degree of 5. In this environment there is either a low enough infection threshold to spread the cascade throughout the network, or there is too high a threshold to infect any early agents.

In SF networks, a wider distribution of degrees allows some cascade spread regardless even with high infection thresholds, but constrains most cascades from spreading across the network. Even with very low infection thresholds, some cascades will start at agents who only have unassailably high-degree hubs as neighbors, preventing any cascade spread (visualized as a green line in Figures 4.1e to 4.1h). Many agents are likely to be connected to

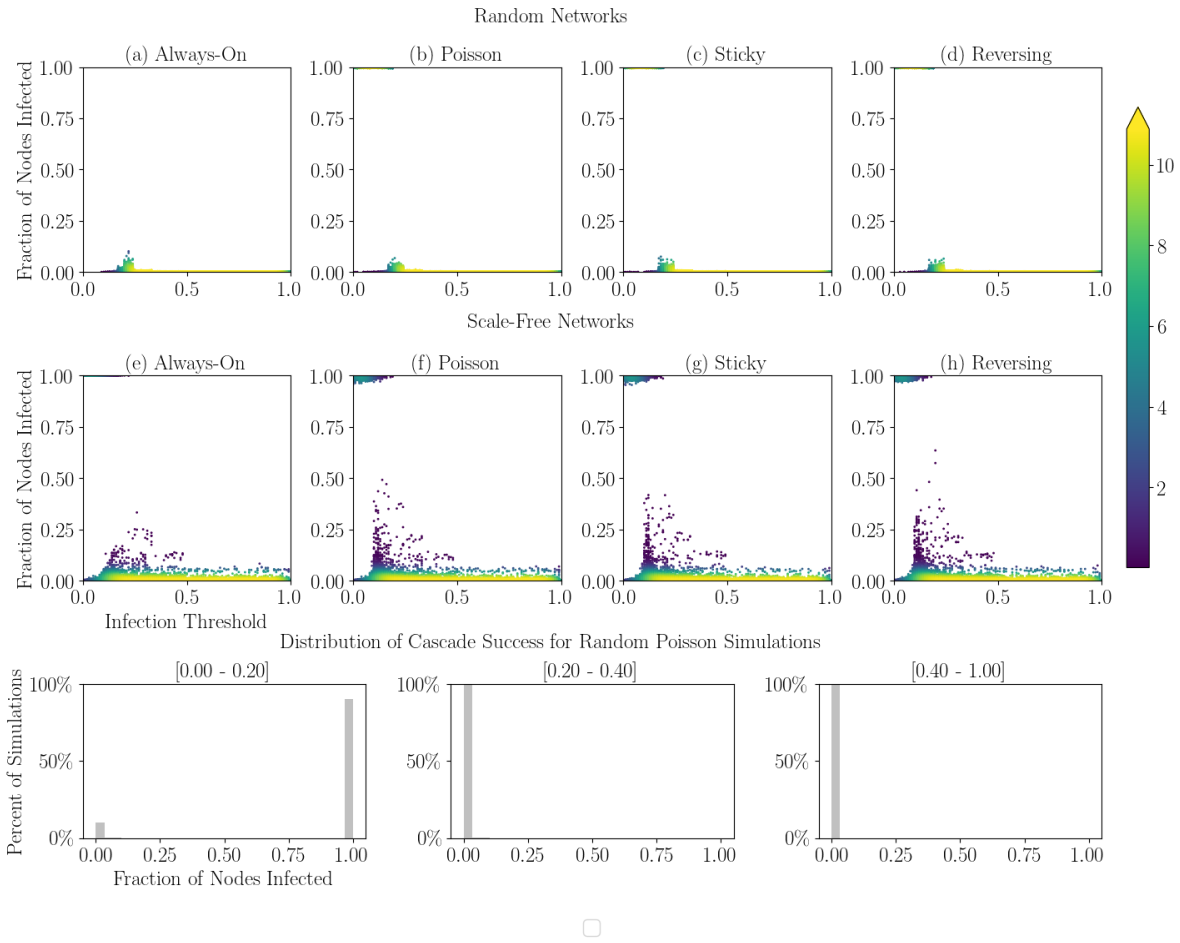


Figure 4.1: Fixed homogeneous infection threshold with Erdős-Rényi (ER) and scale-free (SF) networks. Brighter and more yellow colors indicate more simulations in an end-state, with results to color shown on the color-bar to the right. With random networks there are almost boolean results: cascades either spread through the entire network, or fail almost immediately. A minority of cascades fail even under favorable infection thresholds due to poor starting position of the initially infected agent. With SF networks the range of sometimes-successful inputs increases, because SF networks are much more prone to cascades if the starting infected agent is a hub, which infrequently occurs by chance. In addition, scale-free networks feature a large number of partially successful cascades, capturing 10% or more of the network, but failing to pass 50%. This is also a result of hub behavior, where a cascade may be able to spread through an entire sub-region of the network with vulnerable low-degree agents, but is effectively "quarantined" by an SF hub with too high a degree to infect. Histograms at the bottom of the figure highlight distribution of results within a region of infection thresholds.

some low-degree neighbors, and may spread cascades to these highly-susceptible peers even with an infection threshold higher than the mean degree of the network.

For AlwaysOn agents (Figures 4.1a and 4.1e), cascade success is almost perfectly boolean: Cascades either sweep the entire network, or fail to spread immediately. Success is less definite with other activity patterns, and some cascades halt at 95 or even 90% spread. This is because a cascade can only spread if two peers are active simultaneously, and non-deterministic activity patterns open the possibility that there will never be sufficient active neighbors at any given moment to overwhelm a particular agent. Without inactive periods there can be no activity time inhibitor on when a cascade spreads.

4.2 Information Spread in Human Interaction Networks with Heterogeneous Infection Thresholds

When agents have heterogeneous infection thresholds, the differences between activity patterns and network topologies become more pronounced. When examining the percent of infected agents over a range of infection thresholds, we see mixed phrase transitions of discontinuous first-order and continuous second order, representing much more interesting behavior than the homogeneous case.

For the Poisson, Sticky, and Reversing activity patterns, the cascade results show a continuous second-order phase transition from one to zero, and a discontinuous first-order phase transition from zero to one. For the Always-On pattern, the peak values jump to zero an infection threshold of 0.65 for ER networks, and 0.71 for SF networks, both of which are much higher than the respective homogeneous cutoffs, 0.22 and 0.17. The point of phase transition for Poisson agents is much lower, at 0.51 for ER networks, and just under 0.48 for SF. This discrepancy is the result of the difficulty in building momentum without vulnerable neighbors. Since there are far more vulnerable nodes in a heterogeneous AlwaysOn activation network, cascades will almost always spread to some neighbors.

In SF networks, the peak values show a discontinuous phase transition from almost one hundred percent of the network to zero, similar to the homogeneous case. However, we find that the boundaries between successfully propagating cascade conditions and unsuccessful ones are no longer nearly as tightly defined, and the ending states of the simulations (represented by the grey dots in Figures 4.2e to 4.2h) are more diverse than the homogeneous

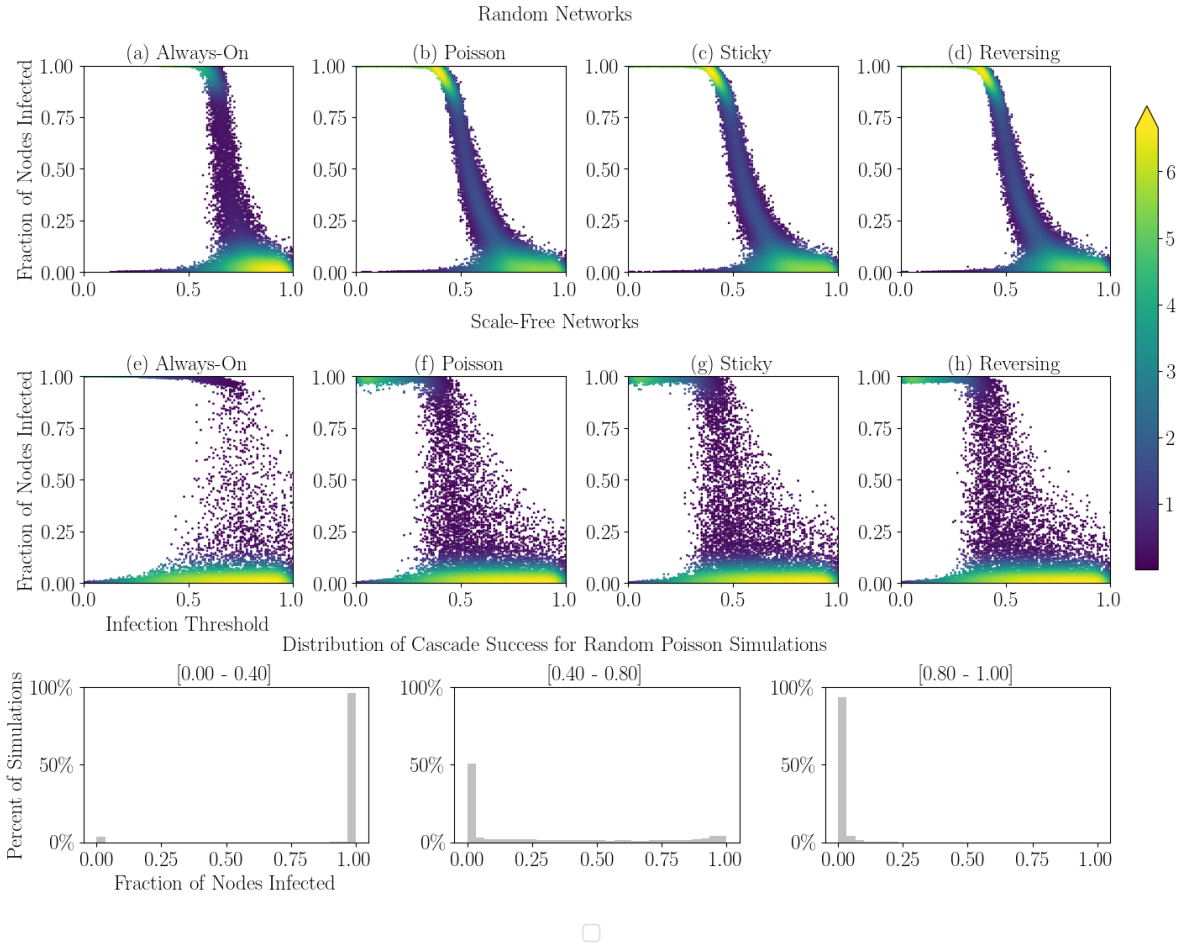


Figure 4.2: Fraction of infected agents after cascades where agents have heterogeneous infection thresholds. (a) For ER networks with Always-On activity patterns, the peak values for each infection threshold jump from a value close to one to zero at a critical point, indicating a well-defined transition from completely successful cascade-spread conditions to complete failure conditions. For networks with other activity patterns (b)-(d), mixed phase transitions appear: one high peak values continuously decrease as the infection threshold increases, and another low peak value disappear after the critical point. In SF networks (e)-(h), regardless of the activity pattern, peak values have discontinuous phase transitions similar to the homogeneous case, but with a much wider range of simulation results (each gray dot).

case. As in the constant infection threshold simulations, AlwaysOn agents have the strictest behavior, with most cascades either succeeding entirely or failing immediately. Other activity patterns yield a variety of intermediate outcomes, with cascades halting everywhere between 0 and 100% success.

The wider phase-transition in SF networks reflects how sensitive these networks are to initial starting conditions: If a cascade starts from a well-connected hub, it can reach many low-degree nodes with ease. Because infection thresholds are heterogeneous, many of these neighbors will be vulnerable. If a cascade starts from a less-connected agent, it will have a small number of neighbors, and some of these neighbors must be vulnerable hubs for the cascade to spread far. Depending on the starting distance from a well-connected hub, and the infection thresholds of all agents between the starting point and that hub, a cascade can halt at many junctions with insufficient momentum.

In addition, we study how the topological characteristics, such as the average degrees in ER networks and the exponent of degree distribution in SF networks, influence the information-spreading process, and find that these topological characteristics have a minimal effect except at extreme values. Very low linking probabilities in ER networks (Figures 4.3a and 4.3c) produce additional vulnerable agents due to their low-degree, and facilitate cascade spread. Cascades can succeed under a variety of conditions depending on the vulnerability of peers, but there is a sharp transition from "successful" to "failed" cascades, and partial-successes can only exist within a very small intermediate parameter-space. In scale-free networks, as shown in Figures 4.3b and 4.3d, this non-boolean space expands dramatically, and many partial-successes, but fewer complete-successes exist.

4.3 The Influence of Activity Periods on Information Spread

Figures 4.4 and 4.5 show that agent activity periods can have a deciding effect on cascade success, but that network topology is the dominant factor. When the infection threshold would otherwise allow a cascade to pass, the activity patterns of agents can constrain a cascade and prevent its spread. With homogeneous infection thresholds, especially on random networks, this is a near-instant phase transition (Figure 4.4), as all agents have nearly equal degree and cascade susceptibility, and inactivity is the only thing preventing infection. With heterogeneous infection thresholds (Figure 4.5) the relationship is more linear, as more resilient agents must be active for long stretches of time to maximize susceptibility,

but more susceptible agents need only be active briefly to become infected by any neighbor. For scale-free networks agent activity patterns still act as a constraint on cascade success, but the results are far less consistent, since the cascade more heavily depends on starting location and the vulnerability and activity of high-degree hub nodes.

4.4 Minimum Starting Infection to Guarantee Total Information Spread

In the results above, we examine under what conditions a cascade can spread throughout the network from a single inception point, but we can instead frame the study as "how many agents must begin infected to spread a cascade throughout a network with particular fixed parameters?" This more accurately represents scenarios such as advertisement or propaganda, which rarely begin from a single point in a community. These intentional cascades not only start from several participants but may have carefully chosen the starting participants to maximize the effect. We study both random and targeted methods for selecting initial participants.

We find that all attack strategies succeed with fewer infected starting agents for highly heterogeneous scale-free networks, and that targeting the hubs in these communities is extremely effective. However, this advantage rapidly disappears as the network conditions become even slightly more chaotic, at which point targeted attack strategies are only moderately more effective than random attacks. In communities with high infection thresholds targeted attacks are indistinguishable in effectiveness, or even perform slightly worse, than random attacks.

As shown in Fig. 4.6, targeted attacks for infection thresholds below 0.5 are generally more successful (require a lower starting infection percentage) than random attacks. Targeted attacks also require far fewer starting agents when the network is extremely heterogeneous in degree distribution ($\gamma = 2$). Interestingly, this advantage does not inversely scale linearly with network heterogeneity, but instead rapidly dissipates when the network is even slightly less heterogeneous. Random attacks remain about equally effective across all network topologies, but are marginally more successful with highly-heterogeneous networks because there is a chance that the hubs will be randomly selected, lowering the average number of starting agents needed for cascade success.

There is a slight jump in targeted attack effectiveness at just under $\gamma = 3$, corresponding with a phase transition from scale-free to random networks [28]. After this phase transition, the highest degree nodes are not hubs, as the network is more densely connected and not dependent on these nodes as bridges. Under these conditions, targeted attacks have a more cumulative effect, as nodes are likely to have edges to more high-degree nodes, briefly improving targeted attack performance.

We are particularly interested in the difference between targeted and random attacks as the infection threshold of the network changes. For an infection threshold of 0.3, where 30% of neighbors must be infected for a cascade to spread, targeted attacks require only about 5% of the network be infected for cascade success, while random networks require closer to 20%. When the infection threshold increases to 40%, this distinction becomes even more dramatic, and targeted attacks require between 20 and 30% starting infection, while random attacks require at least 40%. However, as the infection threshold further increases to 50%, these differences disappear, and both attack strategies require about 70% starting infection unless the network is highly heterogeneous in degree distribution.

Under high infection thresholds, targeted attacks perform marginally *worse* than random attacks for all networks with $\gamma > 2.4$, although the results are extremely dependent on network topology, as seen in the distribution of targeted attack results. In networks where more than half of an agent's neighbors are of high-degree, targeting high-degree nodes is effective. In any network where an agent has few high-degree and mostly *low* degree neighbors, starting from several random positions is more likely to encompass half an agent's neighbors.

These results are intuitive: with a sufficiently low infection threshold, the cascade starting point is less relevant, because it will be able to spread through the network with ease regardless of degree. When the infection threshold is sufficiently high, the starting location is also irrelevant, because a dominating percentage of the network must start infected to spread to any new agents. Only in a middle region, where it is challenging but not overwhelmingly difficult for a cascade to spread, are targeted attacks a dramatic improvement over random ones.

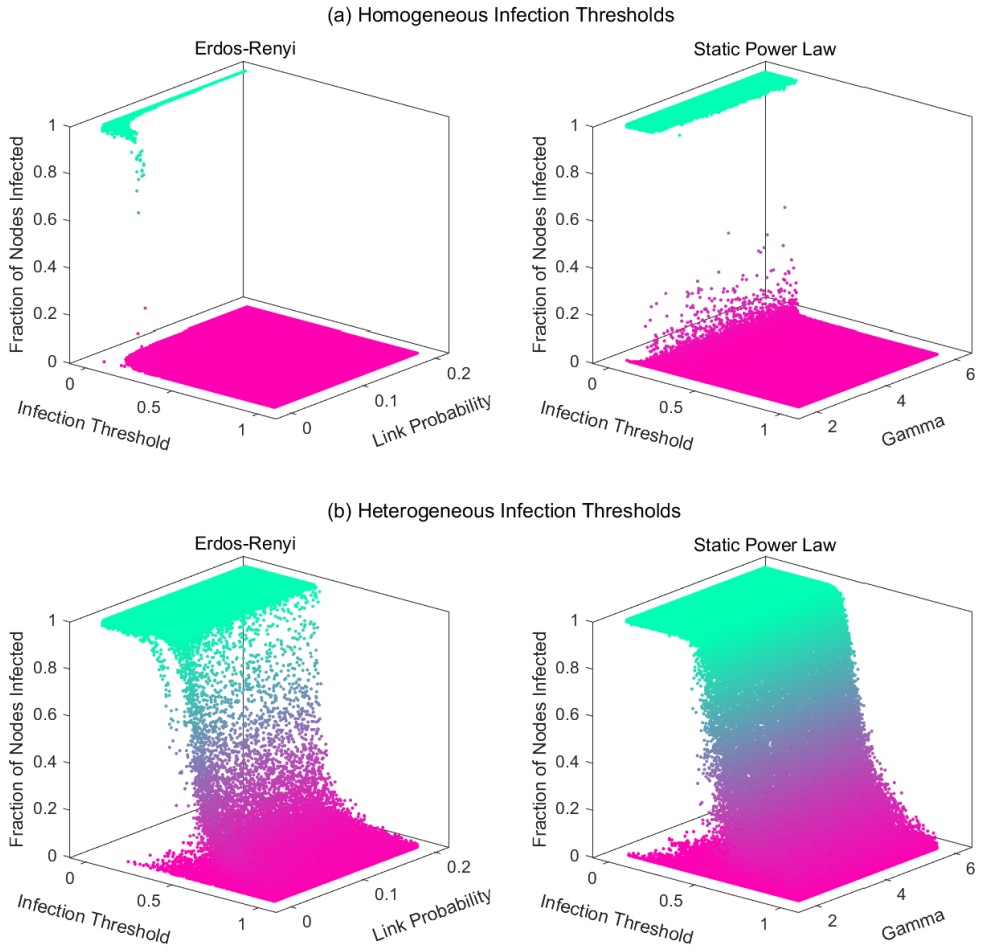


Figure 4.3: Variation on infection threshold and network density with homogeneous and heterogeneous infection thresholds for Poisson agents. Color indicates fraction of nodes infected. Homogeneous networks show near-boolean cascade success, with some partial successes in SF networks, and a small number of nearly-complete-successes for low-link Erdős-Rényi networks. Heterogeneous networks have dramatically more partial-cascade-successes, representing a smooth phase transition from successful to failed cascade-spread conditions. These results reinforce the patterns seen in Figures 4.1 and 4.2, and demonstrate that the distribution of infection thresholds has a far more decisive role in cascade success than network topology. We have excluded all Erdős-Rényi data points with a link probability of exactly zero, as any network with no edges will consist of a single node, and all cascades will by definition spread through 100% of the network.

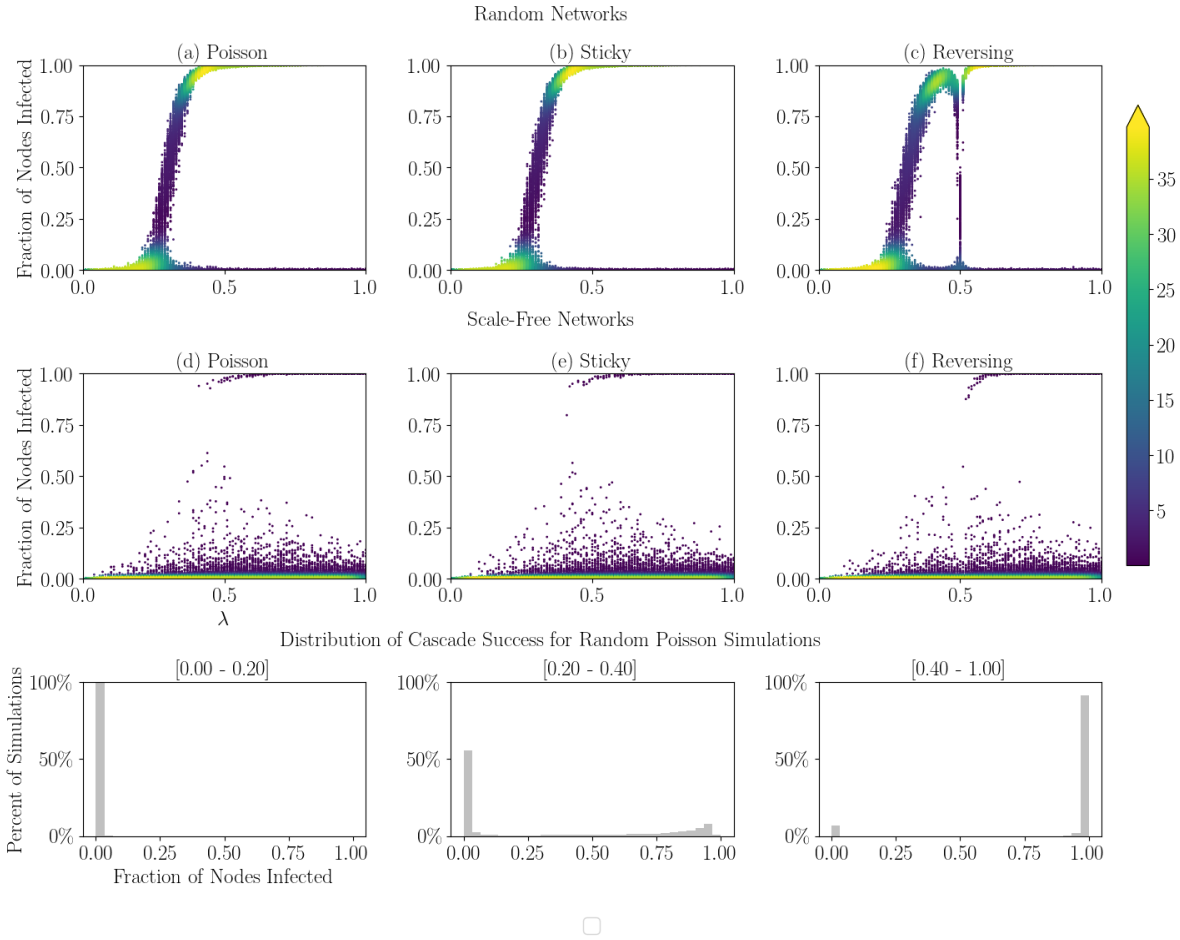


Figure 4.4: Sweep of activity time for infection threshold 0.15, with activity periods from 0% to 100% and homogeneous infection thresholds. In random networks activity patterns can be a deciding factor in cascade propagation, and there is a linear relationship between agent activity and cascade success from about 30% to 50% activity. The anomalous behavior for reversing agents at 50% activity is an unstable equilibrium case where agents are active on either even or odd timesteps, effectively partitioning the network in half and severely hindering the spread of any cascade. With even slight perturbations in agent activity time the agents will naturally "desynchronize" and spread cascades to one another. In scale-free networks the activity time of agents has a less pronounced, but still significant, influence on cascade success. This is likely because the activity time of hubs is significant and constraints cascades, but the majority of agents are of lesser consequence.

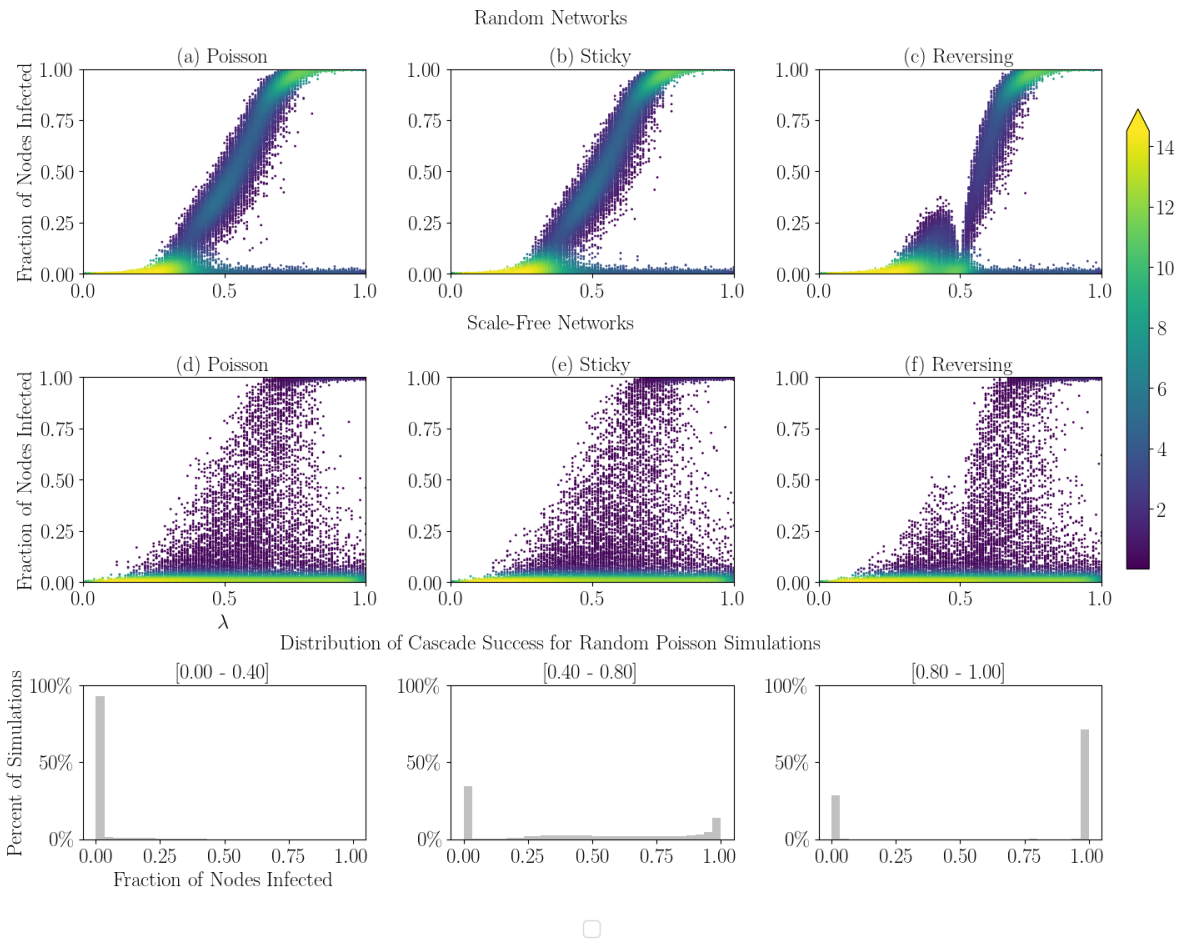


Figure 4.5: Sweep of activity time for infection threshold 0.5, activity periods from 0% to 100%. Heterogeneous infection thresholds. The linear relationship for random networks indicates that activity time is a deciding factor when the network topology and infection thresholds are at an unstable point. As activity moves from 40% to 100%, cascades become more and more successful, until the majority of cascades capture the entire network. As in Figure 4.4, reversing agents demonstrate anomalous behavior with an activity period of exactly 50% because of unstable partitioning among agents active on even and odd timesteps. For SF networks the agent activity period is not as deciding a factor, but is still an enabling one. While most cascades fail regardless of activity period, the number of cascades that partially or completely succeed is bound by the activation period.

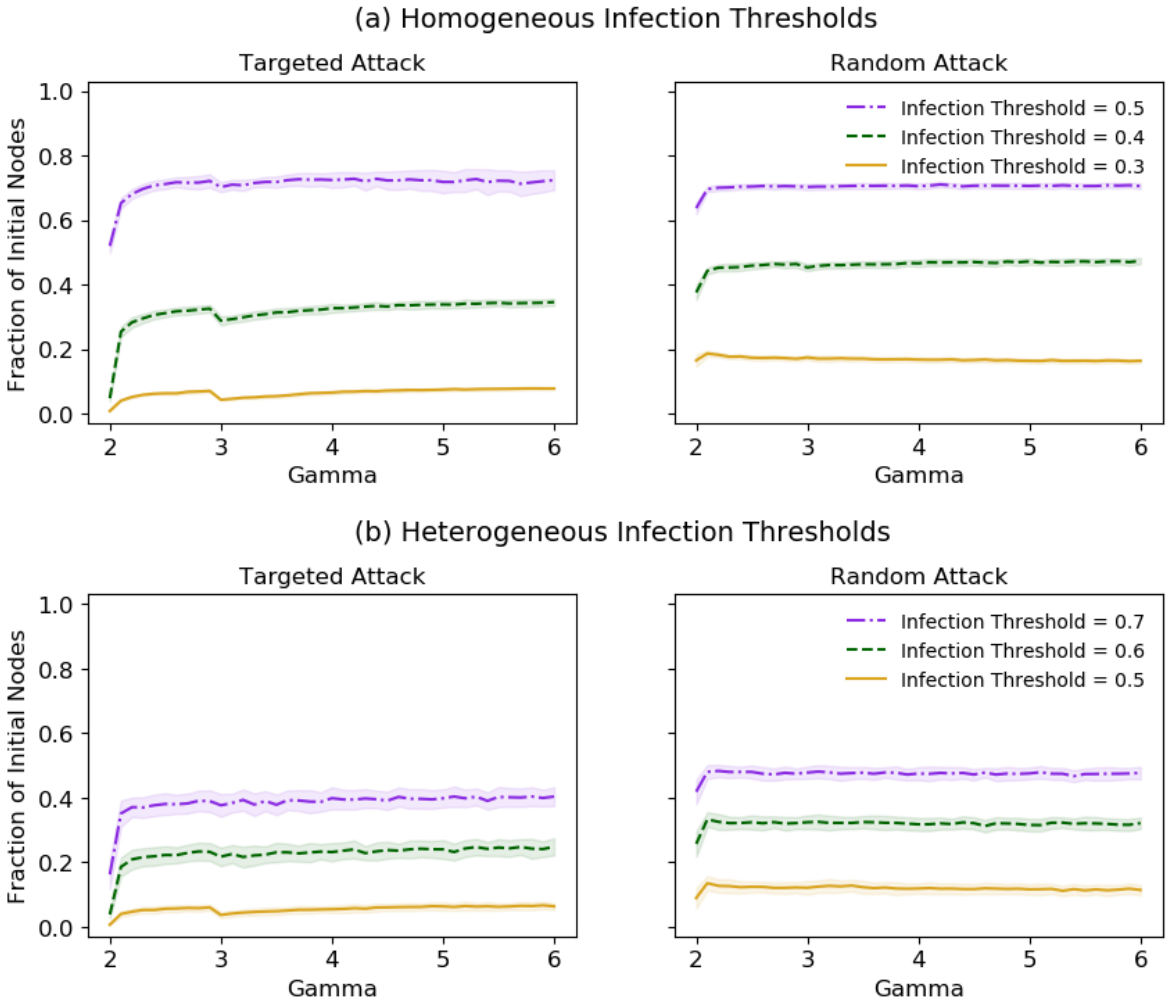


Figure 4.6: **Minimum starting infection needed for cascade success (spread through $\geq 95\%$ of the network). Note that as heterogeneity of SF networks decreases, targeted attacks rapidly become less effective. Networks with more resilient agents are immune to all but the most overwhelming conditions, where cascades start with $60 - 70\%$ of the network already infected. High-infection threshold targeted attacks have a wider range of success conditions than equivalent random attacks (represented by the confidence interval around each simulation line), because the effectiveness of targeted attacks is more dependant on network topology, and therefore less consistent across multiple simulations. Networks with heterogeneous agent susceptibility have similar behavior to homogeneous networks, but are dramatically more vulnerable to attack. Only with extremely high infection thresholds can a heterogeneous network sustain an attack when over 10% of its agents are infected.**

4.5 Real-World Network Trials

Our simulations on real-world communities, shown in Figures 4.7 and 4.8, reinforce our results from artificially-created networks and provide new insights. The citation network behaves almost identically to our generated networks, with a homogeneous phase transition at a slightly lower infection threshold of about 16%, corresponding with a higher average degree of 6.456 than our generated networks that had an average degree of 5.

The Facebook network roughly follows the predicted pattern, but with two distinct differences: The homogeneous phase transition occurs at a much higher infection threshold than the average degree would suggest, and the heterogeneous simulation contains a large number of "plateaus", where a cascades with a wide number of infection thresholds halt.

The first difference can be explained by the distribution of degrees. In random networks all nodes have very similar degrees, so the average degree coupled with a homogeneous infection threshold provides a decent estimate of cascade success. In our artificial scale-free networks there are hubs with a much higher degree close to 20, but with a median of 3 and a mean of 4, the average degree continues to be a rough estimate of cascade success, if less meaningful than with random networks. However, in the Facebook network the degree distribution is dramatically skewed, with hubs holding over a thousand edges, and a median of 25. With such disproportionate influences, the average degree of 44 provides a poor estimate of homogeneous cascade success.

The "plateau" effect is a consequence of the clustering within the network. Large communities consisting of hundreds or thousands of nodes are tightly linked internally, but contain only a handful of bridges to other communities. Cascades "lurch" through sub-communities if they capture sufficient bridges to enter. In the heterogeneous simulation, this results in the staggered cascade result lines representing each large sub-community a cascade was able to capture. We have confirmed this explanation by rendering a sample of simulations with end-states in the plateau region, and include examples in Table 4.1.

We can also eliminate the plateau effect through network perturbations. By adding 1000 random edges we only slightly change the network metrics, but re-integrate the sub-communities, and yield results similar to our simulated random networks, shown in Figure 4.9. The augmented network has an average degree of 44.176 and clustering exponent of 2.516, compared to the original 43.691 and 2.510. This perturbation has a basis in real-world techniques for altering network topology to increase the spread of information [29]. While

the plateau effect does not correspond to a mathematical-theoretic scale-free model, recent research suggests that most real-world social networks are not strongly scale-free, but are weakly scale-free and display similar but not identical behavior to structural models [30].

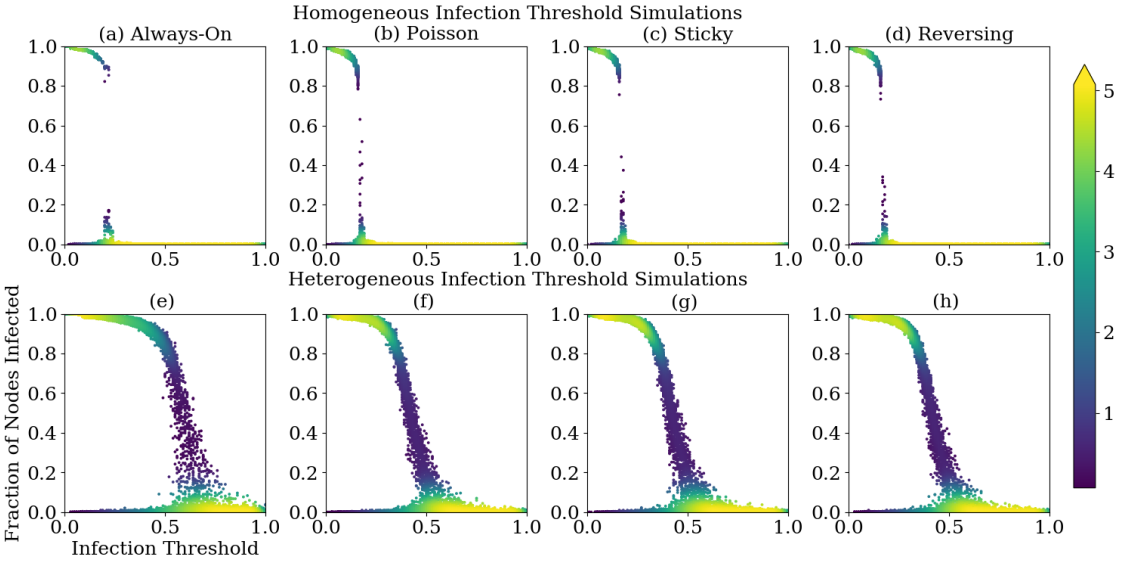


Figure 4.7: Collaboration network for ArXiV papers from the General Relativity and Cosmology category [23]. This real-world scale-free network closely follows predictions from simulated networks in Figure 4.2.

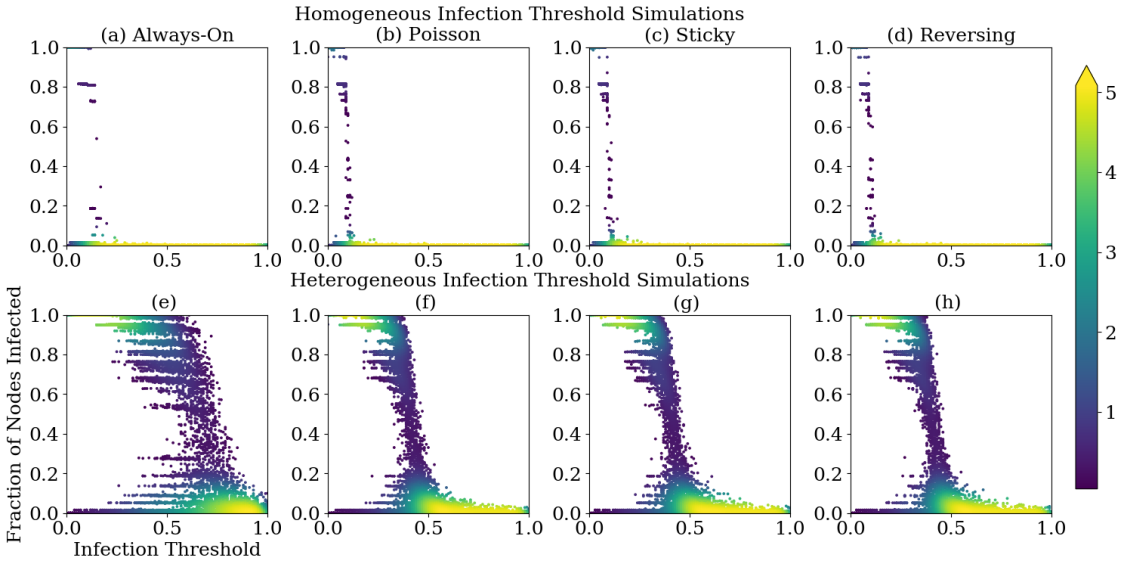


Figure 4.8: Anonymized friendship graph from Facebook data [24]. Cascades generally follow scale-free predictions, but with large visible "plateaus" corresponding to highly isolated friend circles. Most cascades end with some integer number of friend circles captured, with success bounded by the number of bridges they capture.

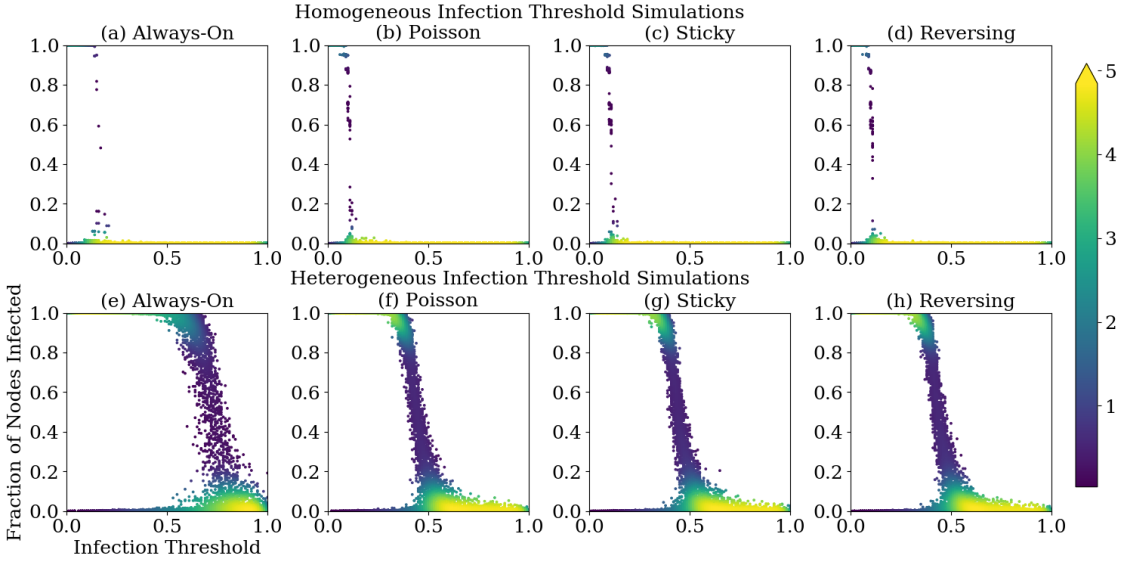
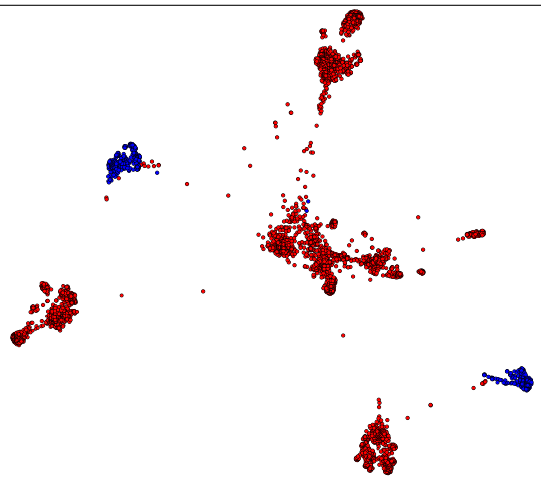
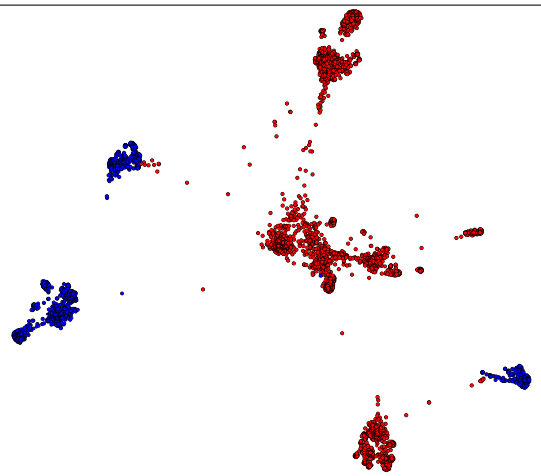
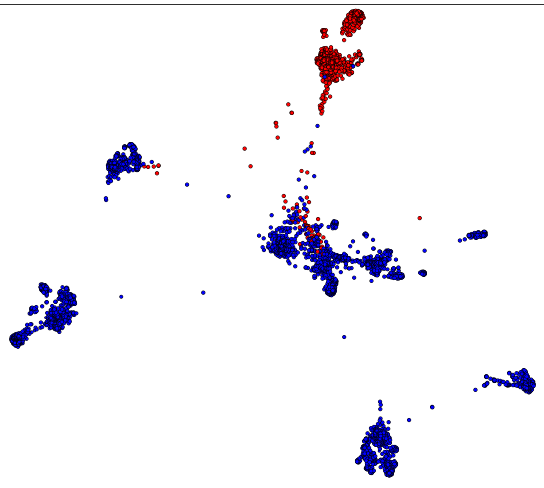


Figure 4.9: Cascade simulations results on an augmented Facebook friendship graph. We have added 1000 random edges (about a 1% increase), which raises the number of bridges between previously isolated sub-communities. The results are quite similar to the artificial random networks in Figure 4.2.

Table 4.1: Renders of simulation results for SNAP Facebook Friend network with varying infection thresholds. Red nodes are infected, blue are uninfected. These results demonstrate the "plateau" effect discussed in Figure 4.8.

Infection Threshold	Network Captured	Snapshot
26%	87%	
32%	68%	
36%	21%	

CHAPTER 5
DISCUSSION

5.1 Conclusion

We outline the relationship between network topology, agent behavior, and cascade viability in human communication networks. We find that scale-free networks allow fewer completely successful information cascades, but many more partially-successful ones. We further find that communities with asynchronously active participants are more resistant to cascades, and that communities with mixed susceptibility to their peers are more vulnerable. Agent activity periods play a significant role in enabling cascades, but cascade success is largely determined by network topology and infection thresholds. Finally, we find that cascades starting from well-connected hubs *are* more successful than cascades starting from random locations in the network, but the differences are only dramatic in highly degree-heterogeneous communities, or communities with a narrow-range of susceptibility to peer influence. Simulations on real-world networks reinforce these results, and illustrate that sub-community dynamics not seen in artificial networks have an inhibiting effect on cascade spread, belaboring the importance of bridges in information spread. These results provide insight into information-spreading mechanisms that could someday aid design of platforms or community structures to facilitate the spread of positive ideas or halt the dissemination of negative content like hatespeech.

5.2 Future Work

The work presented here provides ample opportunity for further study. For instance, our current simulations include only one type of agent behavioral pattern at a time, while one could use heterogeneous collections of activity types with varying activity time and infection thresholds. We studied random and degree-based initial infection schemes, but additional selection criteria are possible, such as "localized" infection [31], [32] by infecting an arbitrary agent and its neighbors. This guarantees a level of "momentum" for the cascade, and may be much more successful than starting at only hubs, particularly in disassociative networks.

This study could be further extended by reproducing research on reinforced agents

from Yuan, Hu, Stanley, *et al.* [33]. Since our model is focused on human societies with heterogeneous activity patterns, it would be valuable to understand whether reinforcing sub-communities is more or less effective in this context. This is particularly interesting in highly scale-free communities, where reinforcing hubs and bridges is most likely to be successful.

Finally, our work assumes each agent has an equal influence on its peers, and information spread is limited solely by susceptibility to peer pressure and the number of neighbors. Allowing heterogeneous influence, where some agents have a disproportionate effect on their neighbors, or even a disproportion effect on a sub-group of neighbors, may produce more realistic results. This disproportionate influence is similar to structuring networks with trait-based homophily, as in Jackson and López-Pintado [34], and could represent one group's trust of a media source. Our Facebook Friend network sample (Section 4.5) includes tight clusters with limited bridges in its network topology, providing an intriguing starting point for exploring in-group and out-group trust. This can ultimately be extended to study the differences in information flow in different types of communities, as defined by social frameworks like group-grid theory [35], which categorizes based on attitudes towards individualism, social hierarchy, and insularity.

REFERENCES

[1] M. Gladwell, *The Tipping Point: How Little Things Can Make a Big Difference*. Boston, MA: Little, Brown, 2006.

[2] P. Crucitti, V. Latora, and M. Marchiori, "A topological analysis of the Italian electric power grid," *Physica A: Statist. Mech. and its Appl.*, vol. 338, no. 1-2, pp. 92–97, Jul. 2004.

[3] R. T. Zhou and R. N. Lai, "Herding and information based trading," *J. of Empirical Finance*, vol. 16, no. 3, pp. 388–393, Jun. 2009.

[4] R. Pastor-Satorras and A. Vespignani, "Epidemic spreading in scale-free networks," *Phys. Rev. Lett.*, vol. 86, no. 14, p. 3200, Apr. 2001.

[5] D. S. Callaway, M. E. Newman, S. H. Strogatz, and D. J. Watts, "Network robustness and fragility: Percolation on random graphs," *Phys. Rev. Lett.*, vol. 85, no. 25, p. 5468, Dec. 2000.

[6] B. Shulgin, L. Stone, and Z. Agur, "Pulse vaccination strategy in the sir epidemic model," *Bull. of Math. Biol.*, vol. 60, no. 6, pp. 1123–1148, Nov. 1998.

[7] K. Saito, R. Nakano, and M. Kimura, "Prediction of information diffusion probabilities for independent cascade model," in *Int. Conf. on Knowl.-Based and Intell. Inf. and Eng. Syst.*, Sep. 2008, pp. 67–75.

[8] P. Crucitti, V. Latora, M. Marchiori, and A. Rapisarda, "Error and attack tolerance of complex networks," *Physica A: Statist. Mech. and its Appl.*, vol. 340, no. 1-3, pp. 388–394, Sep. 2004.

[9] D. J. Watts, "A simple model of global cascades on random networks," *Proc. of the Nat. Acad. of Sci.*, vol. 99, no. 9, pp. 5766–5771, Apr. 2002.

[10] M. Akbarpour and M. O. Jackson, "Diffusion in networks and the virtue of burstiness," *Proc. of the Nat. Acad. of Sci.*, vol. 115, no. 30, E6996–E7004, Jul. 2018.

[11] J. Li, J. Wang, S. Sun, and C. Xia, "Cascading crashes induced by the individual heterogeneity in complex networks," *Appl. Math. and Computation*, vol. 323, pp. 182–192, Apr. 2018.

[12] P. Holme and J. Saramäki, "Temporal networks," *Phys. Rep.*, vol. 519, no. 3, pp. 97–125, Apr. 2012.

[13] I. V. Belykh, V. N. Belykh, and M. Hasler, "Blinking model and synchronization in small-world networks with a time-varying coupling," *Physica D: Nonlinear Phenomena*, vol. 195, no. 1-2, pp. 188–206, Aug. 2004.

[14] M. Gomez-Rodriguez, J. Leskovec, and B. Schölkopf, "Modeling information propagation with survival theory," in *Int. Conf. on Mach. Learn.*, Feb. 2013, pp. 666–674.

[15] M. Squire, "Understanding gray networks using social media trace data," in *Int. Conf. on Soc. Inform.*, Nov. 2019, pp. 202–217.

[16] P. Sobkowicz, "Effect of leader's strategy on opinion formation in networked societies with local interactions," *Int. J. of Modern Phys. C*, vol. 21, no. 06, pp. 839–852, Jun. 2010.

[17] ——, "Quantitative agent based model of opinion dynamics: Polish elections of 2015," *PloS One*, vol. 11, no. 5, p. 0 155 098, May 2016.

[18] S. Galam, "Modeling the forming of public opinion: An approach from sociophysics," *Global Econ. and Manage. Rev.*, vol. 18, no. 1, pp. 2–11, Jan. 2013.

[19] J. Borge-Holthoefer, N. Perra, B. Gonçalves, S. González-Bailón, A. Arenas, Y. Moreno, and A. Vespignani, "The dynamics of information-driven coordination phenomena: A transfer entropy analysis," *Sci. Adv.*, vol. 2, no. 4, e1501158, Apr. 2016.

[20] P. Erdos and A. Rényi, "On the evolution of random graphs," *Publ. Math. Inst. Hung. Acad. Sci*, vol. 5, no. 1, pp. 17–60, Jan. 1960.

[21] R. Albert and A.-L. Barabási, "Statistical mechanics of complex networks," *Rev. of Modern Phys.*, vol. 74, no. 1, p. 47, Jan. 2002.

[22] E. Zio and G. Sansavini, "Modeling failure cascades in networks systems due to distributed random disturbances and targeted intentional attacks," in *Proc. of the Eur. Saf. and Rel. Conf. (ESREL 2008)*, Sep. 2008, pp. 1861–1866.

[23] J. Leskovec, J. Kleinberg, and C. Faloutsos, "Graph evolution: Densification and shrinking diameters," *ACM Trans. on Knowl. Discovery from Data (TKDD)*, vol. 1, no. 1, 2–es, Mar. 2007.

[24] J. Leskovec and J. J. Mcauley, "Learning to discover social circles in ego networks," in *Adv. in Neural Inf. Process. Syst.*, 2012, pp. 539–547.

[25] J. Leskovec and A. Krevl, *SNAP Datasets: Stanford large network dataset collection*, Jun. 2014, [Online] Available: `http://snap.stanford.edu/data` (Accessed Feb. 2, 2020).

[26] F. J. Massey Jr, "The Kolmogorov-Smirnov test for goodness of fit," *J. of the Amer. Statist. Assoc.*, vol. 46, no. 253, pp. 68–78, Mar. 1951.

[27] P. Shannon, A. Markiel, O. Ozier, N. S. Baliga, J. T. Wang, D. Ramage, N. Amin, B. Schwikowski, and T. Ideker, "Cytoscape: A software environment for integrated models of biomolecular interaction networks," *Genome Res.*, vol. 13, no. 11, pp. 2498–2504, Nov. 2003.

[28] A.-L. Barabási *et al.*, *Network Science*. Cambridge, United Kingdom: Cambridge University Press, 2016.

[29] T. W. Valente, "Network interventions," *Science*, vol. 337, no. 6090, pp. 49–53, Jul. 2012.

[30] A. D. Broido and A. Clauset, "Scale-free networks are rare," *Nature Commun.*, vol. 10, no. 1, pp. 1–10, Mar. 2019.

[31] M. M. Danziger, L. M. Shekhtman, Y. Berezin, and S. Havlin, "The effect of spatiality on multiplex networks," *EPL (Europhysics Lett.)*, vol. 115, no. 3, p. 36 002, Sep. 2016.

[32] B. Gross, D. Vaknin, M. M. Danziger, and S. Havlin, "Multi-universality and localized attacks in spatially embedded networks," in *Proc. of the Asia-Pacific Econophysics Conf. 2016—Big Data Anal. and Model. toward Super Smart Society—(APEC-SSS2016)*, 2017, p. 011 002.

[33] X. Yuan, Y. Hu, H. E. Stanley, and S. Havlin, "Eradicating catastrophic collapse in interdependent networks via reinforced nodes," *Proc. of the Nat. Acad. of Sci.*, p. 201 621 369, Mar. 2017.

[34] M. O. Jackson and D. López-Pintado, "Diffusion and contagion in networks with heterogeneous agents and homophily," *Netw. Sci.*, vol. 1, no. 1, pp. 49–67, Apr. 2013.

[35] M. Douglas, "Introduction to grid/group analysis," in *Essays in the Sociology of Perception*, New York, NY: Routledge, 1982, pp. 1–8.