2024

# Models of Information Diffusion and The Role of Influence

Chathura JJ Don Dimungu Arachchige
*University of Central Florida*

# MODELS OF INFORMATION DIFFUSION AND THE ROLE OF INFLUENCE

by

CHATHURA JEEWAKA JAYALATH DON DIMUNGU ARACHCHIGE
B.Sc. University of Colombo, 2014

A dissertation submitted in partial fulfilment of the requirements
for the degree of Doctor of Philosophy
in the Department of Industrial Engineering and Management Systems
in the College of Engineering and Computer Science
at the University of Central Florida

Spring Term
2024

Major Professor: Ivan Garibay

# ABSTRACT

Information diffusion is significant in fields such as propagation prediction and influence maximization, with applications in viral marketing and rumor control. Despite conceptual differences, existing diffusion models may not represent identical underlying generative structures. A classification of diffusion of information models is developed based on infection requirements and stochasticity. The study involves analyzing seven existing DOI models on directed scale-free networks. The distinctive properties of each model are identified through simulations and analysis of experimental results. Our analysis reveals that similarity in conceptual design does not imply similarity in behavior concerning speed, the final state of nodes and edges, and sensitivity to parameters. Therefore, we highlight the importance of considering the unique behavioral characteristics of each model when selecting a suitable information diffusion model for a particular application. We further investigate how the network structure and clustering affect the diffusion of information. Our findings reveal that clustering does not consistently accelerate the spread of information. Instead, the extent to which clustering facilitates the dissemination of information is influenced by the interplay between the specific network structure types and the information diffusion model employed. Another significant aspect of information diffusion is the effect of influential nodes. Identifying highly influential nodes is of great interest for strategic targeting in various applications such as viral marketing and information campaigns. Our follow-up study aims to identify influential nodes using a transfer entropy-based method. In this work, we use our method to identify influential users in Twitter data and compare the results against other existing methods. Finally, we developed a methodology based on Transfer Entropy to evaluate influence in the context of information diffusion. This methodology demonstrated its superiority in predicting user adoption against retweet-based metrics, marking it as a direct and reliable metric for understanding influential users and information diffusion trends.

To my parents, who instilled in me the value of education and celebrated every step of my journey. This is for you.

# ACKNOWLEDGMENTS

I am very grateful for the unwavering support and guidance my advisor, Dr. Ivan Garibay, provided throughout my journey. His consistent support, whether securing additional funding or accessing essential resources, was instrumental in my success. Dr. Graibay's probing questions were crucial in steering my research, and his genuine care for his students fostered a supportive environment in and out of the lab. I extend my heartfelt gratefulness to my co-advisor, Dr. William Rand, whose mentorship was invaluable throughout my Ph.D. Dr. Rand's invaluable feedback and directions consistently improved my work, extending far beyond the scope of our initial collaboration. I am indebted to both Dr. Garibay and Dr. Rand, not only for their mentorship but also for their friendship. Their introductions to key figures in our field opened doors for collaborative ventures, broadening my horizons and enriching my academic journey. I also want to express my gratitude to the rest of the members of my Ph.D. committee, Dr. Thomas O'Neal, Dr. Luis Rabelo, and Dr. Timothy Kotnour, for their valuable feedback and guidance throughout the process. A special thanks is due to Dr. Ozlem Garibay for her cheerful smile and continuous encouragement and support. I thank my dear friend Dr. Chathika Gunaratne, whose encouragement was pivotal in my decision to pursue my Ph.D. in the US. I am also thankful to my friends and colleagues at the Complex Adaptive Systems Laboratory, especially Dr. Chathurani Seneviratne, Dr. Nisha Baral, Dr. Bruce Miller, and Dr. Jasser Jasser. Special acknowledgment is reserved for my dear friend and collaborator, Xiaoxia Champon, from NCSU. Her relentless push for excellence challenged me to exceed expectations in my dissertation work and personal growth, for which I am very grateful. I'm also thankful to Dr. Dan Eilen and Ms. Hannah Faler for their invaluable support, that was crucial to my academic progress. I extend my gratitude to my housemate and dear friend, Mr. Isuru Kuruwitage, whose friendship has been a source of great strength and support throughout the years. To my numerous friends and family worldwide, whom I couldn't mention individually,

I express my sincere appreciation for your friendship and support. Finally, I express my deepest gratitude to my family: my parents and brother for their boundless love and care and my loving wife for her continuous encouragement and unending belief in me. Their love and support have been the bedrock of my career, inspiring me to strive for excellence every step of the way.

This research work was funded through mainly two DARPA programs, SocialSim program (SocialSim HR001117S0018 (FA8650-18-C-7823)) and MIPS program (MIPS HR00112290104 (PA-21-04-06)).

# TABLE OF CONTENTS

# LIST OF FIGURES

# LIST OF TABLES

# LIST OF ABBREVIATIONS

**ABM** Agent-based model

**DOI** Diffusion of Information

**GAM** Generalized Additive Modeling

**NPV** Net Present Value

**OSN** Online Social Network

# CHAPTER 1: INTRODUCTION

Understanding the diffusion of information in social networks is of great interest due to its application in various problems such as marketing, promoting societal benefits, and preventing the spread of misinformation. The speed of this diffusion has been accelerated by the growth of social media. While the dynamics of information diffusion is affected by factors such as peer influence and authority pressure [2], contemporary information diffusion models identify features such as network structure (e.g. scale-free, small-world), node activation mechanisms (e.g. probability-based, threshold-based), and communication method (e.g. peer-to-peer, broadcast) as factors that affect the dynamics of information diffusion within the simulation domain. Together these modeling factors are representing the natural causes such as peer influence. Further, these properties may exhibit interdependence. For example, the internal likelihood of adoption may increase when a significant portion of the neighborhood comprises adopting neighbors. Conversely, the prevalence of adopting neighbors could be influenced by the internal likelihood of adoption, due to homophily. Whether and how a user decides to actively engage with information on social media, which we call the *activation mechanism*, may differ depending on a multitude of factors, including the neighborhood network structure, activation state of neighbors, internal likelihood of adoption, and fraction of adopting neighbors, to name a few.

Classical models of information diffusion [5, 37, 21, 16, 22, 11, 9] utilize a wide variety of node activation mechanisms, yet have not been compared under a common framework. The lack of such an analysis has led to the danger of misinterpreting a well fit model as confirmation that the underlying conceptual design inspiring this model is the real-world explanation of the human behavior that generated the observed information cascade. Hence, we explore the importance of understanding whether or not models with differing conceptual explanations have the same coverage of the simulation landscape, and if so, the conditions under which such similarity, and

ambiguity in explanation, occurs. Information is propagated through social networks in a way such that the person who adopts some received information attempts to share that information with their peers. Classical models of information diffusion explain this adoption behavior through a variety of conceptual definitions including adoption as *infection*, threshold-based activation, social norm-driven activation, etc. These conceptual definitions result in varying model mechanisms such as, Bayesian likelihood of adoption, adoption based on the number of known adopting neighbors, and adoption based on fraction of adopting neighbors.

The first study contributes a conceptual framework that classifies models of information diffusion into four classes based on *neighbor knowledge* and *stochasticity*. We perform comparisons between existing models that fall under these four conceptual classes in order to establish the simulation conditions under which these classes have similar final infected ratios, given their inputs. In particular, we evaluate the Linear Absolute Threshold Model (LATM) [22], the Linear Fractional Threshold Model (LFTM) [22, 51], the Independent Cascade Model (ICM) [21], the Bass-Rand-Rust Model (BRRM) [37, 38], the Stochastic Linear Fractional Threshold Model (SLFTM) [6], the Stochastic Linear Absolute Threshold Model (SLATM) (adopted from [6]), and the Dodds-Watts model (DWM) [17] on directed scale-free networks under similar model parameter values. Our results demonstrate that, despite belonging to the same conceptual class, the outcomes of models may be completely different from each other. Furthermore, our results confirm that despite belonging to conceptually different classes, models may produce equal outcomes. In essence, we conclude that similarities (or differences) in conceptual design does not guarantee similar (or different) simulation outcomes.

As we shown that the choice of DOI model affects the final state of the system, it raises caution when extrapolating conclusions drawn with one DOI model into another. While the choice of DOI model is found to be a critical element that affects the diffusion outcome, it is also important to investigate the significance of network structure and other network properties in determining the

outcome. In addition, the effect of clustering on DOI have been discussed in previous work[52, 10] and Watts and Strogatz[52] showed the efficiency of simple contagion models on highly clustered small-world networks. Further, Centola, Eguíluz, and Macy[10] have investigated the critical thresholds of threshold based complex contagion models on different types of networks that have different clustering. However, there is a lack of understanding to how clustering in a network affects different DOI models under different network types. In particular, a question arises as to whether clustering holds greater significance than network structure when assessing the outcome of a DOI model. In order to study this phenomenon, we investigate the effect of clustering on the diffusion across multiple network types and DOI models. The following classical network types are investigated in this study: Random networks (R) proposed by Erdos and Rényi [19], Small-world networks (SW) proposed by Watts and Strogatz [52], and Scale-free networks (SF) proposed by Barabási and Albert [4]. With the knowledge gained from the first study, we chose only the three main classical DOI models to be investigated under this study: Independent Cascade Model (ICM) [21], Linear Absolute Threshold Model (LATM) [11, 22], and Linear Fractional Threshold Model (LFTM) [22, 51]. Through this study we show that the effect of clustering coefficient on the information spread is dependent on network type, network parameter, DOI model, and DOI model parameter.

Thirdly, we look at a novel methodology for identifying influential nodes within the information diffusion space. While we have demonstrated an understanding of the differences between information diffusion models and their simulation outcomes with respect to simulation parameters, another key aspect that affects information diffusion is the influential actors (nodes). Influential nodes in online social networks are target nodes of interest that are considered important. The interest might be based on some factor such as generation of large number of retweets and popularity for the content, with the intention of making it viral. There have been many studies that employ basic measures such as number of retweets, number of followers, and centrality of the user in the

network structure for studying influential nodes [12, 3, 27, 29]. We propose a Transfer Entropy-based method to measure a node's influence on information diffusion over a given scenario. We then demonstrate that this measure can estimate user adoption. By comparing this measure against the retweet-based measure using the same methodology, we show that the Transfer Entropy-based measure is a unique and valuable indicator of influence.

## Statement of Contributions

The first two parts of is study contributes to the knowledge of dynamics and characteristics of information diffusion. As described above, information diffusion models identify features such as network structure and node activation mechanism as factors that affect the diffusion of information. The first part of this study aims at comparing existing information diffusion models under a common framework. The second part of the study builds upon the initial findings by exploring the impact of clustering on information diffusion across various network structures and in the context of different diffusion models.

- Propose a generalized conceptual framework for information diffusion models by introducing a generalized form for information diffusion models.
- Classify traditional models based on the introduced conceptual framework into four conceptually distinct classes.
- Develop a tool for comparison of information diffusion models under the proposed conceptual framework.
- Provide evidence that model specific parameters of information diffusion models that infects all reachable nodes are only useful in changing the speed of infection.

- Demonstrate that a stochastic version of a model created by appending a probability check to the final step of the existing rule will yield a model that has its final state bounded by the final state of the original model.
- Illustrate that similarity in conceptual design does not imply similarity in behavior concerning speed, final state of nodes and edges, and sensitivity to parameters. Thus, highlight the importance of considering the unique behavioral characteristics of each model when selecting a suitable information diffusion model for a particular application.
- Show that the existence of clustering only sometimes accelerates the spread of information. The interaction of the type of network structure and diffusion model determines how much clustering accelerates the spread of information.

In the third part of the dissertation we explore how information diffusion is affected by peer influence and authority pressure [2]. Therefore, highly influential nodes are of great interest [1]. There are many methods of identifying influential nodes in literature. For example, on Twitter: the number of retweets, number of followers, and the ratio between posts and received retweets are some of these measures [12, 24]. The second part of this study aims to propose a novel measure for identifying influential nodes based on analysis of user activity over time. Expected contributions of the second study are:

- Propose a Transfer Entropy-based method for identifying influential nodes.
- Develop a tool for identifying influential nodes based on the proposed strategy and compare against other existing strategies.
- Increase the understanding of measures and strategies for identifying influential nodes.
- Show that the Transfer Entropy-based method for identifying influential nodes is capable of estimating user adoption.

Statement of Originality

Parts of this work have been published in a journal and parts of this work have been presented at conferences without publishing in procedures. Other than the works presented and discussed in the manuscripts that follows, the rest of this dissertation has not been published publicly at the time of writing.

- Chathura Jayalath, Xiaoxia Champon, William Rand, and Ivan Garibay. Quantifying Influence on Online Social Media Using Transfer Entropy. Manuscript in preparation.
- Chathura Jayalath, Xiaoxia Champon, William Rand, and Ivan Garibay. Quantifying the Effect of Clustering Coefficient on Models of Information Diffusion. Manuscript in preparation.
- Chathura Jayalath, Xiaoxia Champon, William Rand, and Ivan Garibay. The Effects of Clustering Coefficient on Models of Information Diffusion. Presented at International Conference of The Computational Social Science Society of the Americas. 2023.
- Chathura Jayalath, Chathika Gunaratne, William Rand, Chathurani Senevirathna, and Ivan Garibay. A generalization of threshold-based and probability-based models of information diffusion. Advances in Complex Systems, 2023.
- Chathura Jayalath, Chathika Gunaratne, Bill Rand, Chathurani Senevirathna and Ivan Garibay, Final states of Threshold based Complex-contagion model and Independent-cascade model on directed Scale-free Networks Under Homogeneous Conditions, Poster-presentation at Tenth International Conference on Complex Systems, 2020.

# CHAPTER 2: LITERATURE REVIEW

The dynamics of information diffusion are shaped by various factors, including the network's structure, mechanisms of node activation, the nature of the information, methods of communication, and the resolution and scale at which these processes are observed. Moreover, these factors can affect one another; for instance, the way nodes are activated can vary based on their surrounding network structure, and the network structure, independent of the communication method employed, may constrain the extent of communication reach. Therefore, we first look at the models of network structures and then discuss existing literature on information diffusion models.

## Network Models

Contagions have been studied across various types of networks, with different network types identified through the analysis of connectivity properties. Three main models of networks are prominently featured in the literature: random networks [19], small-world networks [52], and scale-free networks [4].

*Random Networks*: In a random network, links between nodes are created purely stochastic, meaning that any two nodes have a constant probability of being connected. This model, extensively studied by Erdős and Rényi in their pioneering work[19], exhibits a Poisson degree distribution. The degree of most of the nodes in random networks is comparable. Therefore, the significant degree differences observed in real networks is absent in random networks [34]. In real networks, a significant number of highly connected hubs exist, which are also absent in random networks. While random networks can serve as a baseline for understanding more complex structures, they are limited in their ability to capture the clustering and community patterns observed in real-world

networks.

*Small-world Networks*: Small-world networks was a network model introduced by Watts and Strogatz [52] influenced by the the Small-world phenomenon [49]. These networks are characterized by their high clustering coefficient and short average path lengths. These networks are in a interpolation between regular lattices and random networks, featuring a tightly knit structure where most nodes can be reached from any other through a small number of steps.

*Scale-free Networks*: Scale-free networks are distinguished by their power-law degree distribution, where a small number of nodes (hubs) have a very high degree, while the majority of nodes have very low degree. Barabási and Albert provided a foundational model for understanding such networks' growth dynamics and robustness [4, 34]. Scale-free networks are common in both the natural and man-made world, and they are found in the structure of the internet, citation networks, and protein interaction networks, among others. Their topology makes them highly robust to random failures but vulnerable to targeted attacks on their hubs [34].

Compared to other network models, scale-free networks more closely resemble real-world networks, making them suitable for simulations aimed at understanding information diffusion in social networks. The presence of hub nodes in scale-free networks allow the study of actors with large following (potentially influential actors) and their impact on the spread of information within the network. Further, scale-free networks follow a power-law distribution in node connectivity which is also a pattern observed in online social networks [4].

## Diffusion of Information Models

Various information diffusion models exist in the literature. Non-linear dynamical models provide the facility to derive analytical solutions for the overall system in different conditions. This makes

it easier to identify properties such as tipping points [47]. Detailed micro-level modeling is impractical with this type of models. In contrast, agent-based models provide flexibility in studying emergent phenomena that arise from individual interactions of agents in the system, which is difficult to capture in differential equation systems. This is especially true in cases where agent behaviors and attributes are heterogeneous, or the interaction topology (network structure) is heterogeneous. Therefore, agent-based models are more useful in modeling systems such as information diffusion where individual behaviours can significantly influence the outcome of the system (for example: a highly connected agent becoming infected with information might flood the whole network with that infection). Moreover, agents that exhibit complex behaviour could easily be modeled within ABMs[18]. In the following we briefly go through various existing models of information diffusion and how they have inspired us in identifying information diffusion mechanisms that are based on both probability-based and threshold-based techniques. The description of the implementation of the models are given in chapter 3.

One of the earliest diffusion models, the Bass model, describes the diffusion of adoption of an innovation and uses a hazard rate model to describe the population-level adoption of innovations [5]. Though the model is a population-level model, it describes a conceptual, individual-level mechanism, where the propagation of adoption was modeled as a consequence of independent decisions and peer-to-peer influence. Early adopting consumers in the Bass model often adopt the innovation independent of the choice of other individuals. Sometimes this is described as having occurred due to advertising or mass media. Later on the adoption decisions of individuals are affected by both their own independent decisions and social pressure. The probabilities of adoption based on independent decisions and social pressures are called the innovation probability and imitation probability, respectively. The Bass model was not really intended as a predictive model, but as a descriptive model that could be fit to empirical data. This model was then turned into an agent-based model by Rand and Rust [38], which we have adopted in this article as BRRM

(Bass-Rand-Rust Model). The BRRM was later used in modeling information diffusion in urgent situations and the authors suggest a range of parameter values that match against information diffusion on social media [37].

In a different diffusion model, Granovetter [22] talks about explaining band-wagon behaviours using a threshold where each individual has a threshold of their neighbors that need to become activated before the focal user becomes activated. Some information diffusion literature have adopted this threshold as a fraction, and defined their rule of infection such that if the fraction of infected neighbors (number of infected neighbors divided by the size of the neighborhood) exceeds this threshold then the focal agent becomes infected [51]. This is the basis of what we call the Linear Fractional Threshold Model (LFTM). This model has been extended to also examine an absolute number of neighbors, as opposed to a fraction [11], which is the basis for the Linear Absolute Threshold Model (LATM). An interesting modification to these linear threshold models was suggested by Bohlmann et al. [6]. They proposed performing a coin toss (testing a probability of 0.5) at each time-step before deciding to infect a node that have satisfied its threshold condition. We implmeneted versions of the Bohlmann et al. model for both the LFTM and the LATM (the original paper only applied this technique on the LFTM).

A cellular-automata based model was proposed by Goldenberg et al. [21] for simulating diffusion of information through advertising and word-of-mouth[21]. This model considers two types of neighbors: strong ties and weak ties. The probability of activation through strong ties is larger than the probability of activation through weak ties and the probability of activation through weak ties is larger than the probability of adopting because of advertising. Goldenberg et al. [21] show that the effect of advertising is superior at the beginning of the diffusion process and the effects of strong ties and weak ties are more significant in the middle and the late stages. It is this model that inspired the model that we refer to as the Independent Cascade Model (ICM).

The model proposed by Dodds and Watts [16] for simulating contagions was able to generalize SI, SIS, SIR, and SEIR compartmental models. Their paper was focused on generalizing these different compartmental models in epidemiology into one probability-based model for analysing their dynamics [16, 17]. However, in this work we focus on generalizing of both threshold-based and probability-based models under the SI compartmental model. We adopted this model into our analysis by including network structure based contact conditions and filtering out only the SI part of it.

## Approaches for measuring influence

Influential nodes in online social networks are target nodes of interest that are considered important. There are many ways to define influence in information diffusion domain. In most of online social networks such as Twitter, users create and post content, vote content, and follow other users [53]. Based on these actions we could define and quantify the influence of a user by counting creations/posts generated, or number of posts/votes/followers received. These counts are the most basic ways of measuring influence. In Twitter context, these would be number of tweets created by the user, number of retweets received, number of followers, and number of times a user is mentioned. Earilest work on literature is comparing these measurements to see how they differ from each other [12, 3]. In their work, Cha et al. [12] and Badashian et al.[3], compared these measurements against each other to show that having many followers doesn't necessarily generate more posts (e.g. retweets). A later work by Qiu et al. [35] used machine learning techniques for analysing follower networks to identify influential users. However, their methodology didn't yield strong results potentially due to the usage of follower network, therefore, confirming the findings of Cha et al. [12] as well.

Another methodology of measuring importance is based on various network centrality measure-

ments such as closeness centrality, betweenness centrality, eigenvector centrality, page rank, and Katz centrality [27, 29]. Some research have attempted using information theory based method as well for measuring influence [48, 50, 42]. Ver Steeg and Galstyan [50] used Transfer Entropy (TE) and showed that it could capture relationships that are not visible when utilizing follower or mention networks. Transfer Entropy, introduced by Schreiber [41], is a information theoretical measurement based on Shannon entropy [43]. Given two random processes, TE quantifies how much uncertainty in predicting the next state of one process is reduced by incorporating the histories of both processes. Senevirathna et al. [42] used TE based influence measurement to identify interactions between different types of influence in information diffusion, as a hierarchical cascade of influence. They showed that there is a a significant difference between the users that are in the top and bottom of the hierarchy.

# CHAPTER 3: A GENERALIZATION OF THRESHOLD-BASED AND PROBABILITY-BASED MODELS OF INFORMATION DIFFUSION

In this chapter, we describe the first study. In the following methodology section we go introduce definitions and the models, and then we go through the process of creating a generalized form of diffusion of information (DOI) models based on our conceptual framework. Lastly, we introduce the details of the conducted experiments and their simulations.

## Methodology

We consider information diffusion as a process involving two agent states, *susceptible* and *infected*, during short-term information cascades where agents do not experience any loss in acquired information.[1] For consistency, we refer to a person who has adopted information themselves due to exposure to adopting neighbors, as an *infected* agent. Infected agents exhibit activity that promotes the further propagation of the adopted information by their neighbors in the network. Furthermore, we define any person, who is not *infected*, as a *susceptible* agent. A susceptible agent who was exposed to a piece of information and yet did not become infected would still be able to become infected in the future.[2]

[1] Similar to class of Susceptible-Infected (SI) model of contagion in epidemiology [15]

[2] For the purpose of this study, we do not consider other states such as *recovered* state, which represents real-world individuals forgetting information over time, as we are more interested in short-term information propagation, where the likelihood of information loss is negligible.

*Definition of models*

Traditionally in diffusion models, there is either a seeded set of nodes, or a probability that nodes can adopt without social influence. Without these mechanisms no adoption would occur since the basis of these models is a social mechanism of adoption that can not occur if there are no adoptions. In this study, to simplify the comparison between the models, we assume that there is a seeded set of nodes which are chosen randomly at the beginning of each simulation. This seeded set of nodes is taken as a parameter for the model as the initial fraction of infection ($\phi_0$).

*Linear Absolute Threshold Model (LATM):* In LATM, a node becomes infected if its threshold is satisfied by the number of incoming edges from its infected neighbors. The main model parameter of LATM is the threshold value ($\psi$) of nodes.

*Stochastic Linear Absolute Threshold Model (SLATM):* In SLATM, a node becomes infected if two conditions are satisfied. Firstly, its threshold must be satisfied by the number of incoming edges from its infected neighbors. Secondly, if the threshold condition is satisfied, then there is a probability ($p_\omega$) that the node becomes infected. This model is adopted from the work by Bohlmann et al. [6]. The main parameters of SLATM are the threshold value ($\psi$) and the probability value ($p_\omega$).

*Linear Fractional Threshold Model (LFTM):* In LFTM, a node becomes infected if its threshold is satisfied by the fraction of infected neighbors. The main model parameter of LFTM is the fractional threshold value, $\theta$, ($0 \leq \theta \leq 1$) of nodes.

*Stochastic Linear Fractional Threshold Model (SLFTM):* Similar to SLATM, in SLFTM, a node becomes infected if two conditions are satisfied. Firstly, its threshold must be satisfied by the fraction infected neighbors. Secondly, if the threshold condition is satisfied, then there is a probability ($p_\omega$) that the node becomes infected. This model is adopted from the work by Bohlmann et al. [6].

The main parameters of SLFTM are the threshold value ($\theta$) and the probability value ($p_\omega$).

*Independent Cascade Model (ICM):* In ICM, nodes are infected through imitation and sometimes through innovation. The model considers imitation as the process of a node becoming infected due to the influence of its infected neighbors. An infected node can only infect its neighbors in the time step immediately after it becomes infected. After that it is assumed that the infection failed, but the neighbor node can become infected by one of its other neighbors that becomes infected at some point in the future. ICM utilizes a probability parameter, $q$, which is the probability that a focal susceptible node becomes infected by a neighboring infected source node. In certain versions of the model, innovation is modeled as a probability of a node becoming infected due to events that are exogenous to the studied system. This innovation probability is a property of the system. In this study, we assume that there is no external effect on the diffusion process to simplify the comparisons against other models such as LATM and LFTM [5, 37, 21].

*Bass-Rand-Rust Model (BRRM):* This model was proposed as an agent-based version of the original Bass model which was used to describe the adoption of consumer durables [5, 38]. Similar to ICM, in BRRM, nodes are infected either through imitation or innovation. In the imitation process, nodes are infected through a probability which is dependent on a base imitation probability, $q_b$, and the fraction of infected neighbors, $f$. The probability of adopting is then $q_b * f$. However, unlike ICM, an infected node in BRRM gets a chances to infect its neighbors at every time-step. In the innovation process, nodes are infected through a probability which defines the innovation probability, which represents the exogenous effect on the diffusion process, similarly to ICM.

*Dodds-Watts Model (DWM):* This model was proposed to be fairly general and accommodates any population of individuals in contact, including a network structure [16, 17]. In DWM, each node keeps track of dosage of exposure that it has received. At each time-step, there is a probability $p_c$ that a susceptible node receives a dose of exposure from an infected neighbor (if there is at least

one infected neighbor) due to contact. If the number of received doses satisfy a given threshold value $k$, then the node becomes infected.

Since our goal is to analyze conceptually distinct models under a common framework, we first formulated the general mechanism of information diffusion models. Using this formulation, we created a conceptual framework based on two properties of information diffusion models: *neighbor knowledge*, i.e., how much local neighborhood information is considered, and *stochasticity*, i.e., whether the model is random at all. These two properties are used to create a two-by-two table, which we used to identify four mechanistically distinct classes of models, which match up with the classical DOI models. In order to compare differences between emergent properties of these conceptually distinct models, we first compared the state space after the model has run to completion (final state) and how soon each model reaches middle and final states, and then evaluated the importance of model parameters relative to network structure and initial condition by comparing their effect on the variance of final state and the variance of propagation speed.

### *Conceptual Framework for Models of Information Diffusion*

When designing information diffusion models, different narratives of how diffusion occurs leads to different conceptual models with different mechanisms. Defining the conceptual space of diffusion models will provide us with a way to explore the space of alternate hypothetical causes.

Classical agent-based models of information diffusion work by making node-level comparisons of the intensity of exposure to infected neighbors $F$ to a threshold function $G$, which we would refer to as the rule of infection $\Lambda$. $F$ and $G$ are functions such that codomain is $\mathbb{R}$ and definition of $\Lambda$ is given by the Eq. 3.1. The models differ, in how $F$ and $G$ are defined. Some examples of classical models of information diffusion and their corresponding $F$ and $G$ definitions are given in Table 3.1.

$$\Lambda \equiv F \geq G \tag{3.1}$$

Table 3.1: Examples of models that use various $F$ and $G$ functions.

| $F$ | $G$ | Example models |
|---|---|---|
| $F = \lvert I_x \rvert$ | $G = c$ | LATM, 1st condition of SLATM |
| $F = \frac{\lvert I_x \rvert}{\lvert N_x \rvert}$ | $G = c$ | LFTM, 1st condition of SLFTM |
| $F = c$ | $G \sim U(0,1)$ | 2nd conditions of SLATM and SLFTM |
| $F = 1-(1-q)^{\lvert I_x \rvert}$ | $G \sim U(0,1)$ | ICM |
| $F = q_b \frac{\lvert I_x \rvert}{\lvert N_x \rvert}$ | $G \sim U(0,1)$ | BRRM |
| $F = \lvert M_x \rvert$ | $G = c$ | DWM |

$I_x$ is the subset of neighbors that are infected.
$N_x$ is the set of neighbors.
$M_x$ is the set of doses stored in memory of the node.
$c$ is a fixed constant for the simulation.
$U(0,1)$ is drawn from a uniform random number generator.

The functions $F$ and $G$ determines the behaviour of the information diffusion model. The function $G$ controls whether the mechanism is stochastic or not and $F$ controls how much neighborhood information is considered.

In formal terms, a key property of $F$ is such that $F$ may be subject to its *neighbor knowledge* of the neighborhood either completely or partially, i.e., either consider the entire neighborhood $N_x$ or a subset of neighbors $A_x$ where $A_x \subseteq N_x$. Moreover, a key property of $G$ is that $G$ may be either deterministic or stochastic, i.e., either a constant such as a model parameter (e.g. $\psi$ of LATM) or a pseudo random number drawn at each instance of execution. Notice that some models such as SLATM and SLFTM use multiple conditions that are applied in a step-wise fashion as shown in the table. Moreover, some models such as DWM, performs calculations that requires maintenance of a memory for each node. The calculation of such memory variables may not be captured inside the $\Lambda$ function although it affects the $\Lambda$ (e.g., Memory of received doses in DWM). We plot these two dimensions $F$ and $G$ in Table 3.2 in order to find different classes of models that we could

generate. Further, we identified an example classical diffusion model for each class and they are given in the table.

With this distinct breakdown of the rule of infection we have identified two distinct and mutually exclusive dimensions which generates the following conceptual framework of DOI models (Table 3.2).

- **Neighbor knowledge of infection requirement**

  Whether the whole local neighborhood of the susceptible node is considered when deciding whether to become infected or not. We classify DOI models in to two classes based on neighbor knowledge of the rule of infection.

  1. **Complete**: The infection mechanism of these models depend on the state of the whole local neighborhood. For example, a model that changes its probability of infection of a susceptible node based on the fraction of infected nodes in its local neighborhood has a complete neighbor knowledge.
  2. **Partial**: The infection mechanism of these models depend only on a selected subset of the local neighborhood

- **Stochasticity of the model**

  Whether the rule of infection is deterministic or stochastic when there is at least one infected neighbor.

  1. **Deterministic**
  2. **Stochastic**

The $\Lambda$ of LATM, LFTM, ICM, and BRRM can be written algebraically without conditionals. Therefore, determining the stochasticity and neighbor knowledge of these models is straight for-

ward. LATM and LFTM are deterministic while BRRM and ICM are stochastic. LATM and ICM requires only partial neighbor knowledge while BRRM and LFTM requires complete knowledge of neighborhood infection. The $\Lambda$ of DWM, SLATM, and SLFTM cannot be written algebraically without conditionals. Also models such as DWM requires maintaining memory.[3] Therefore, determining the stochasticity and neighbor knowledge of these models is complicated. The neighbor knowledge requirement of DWM is partial since it only requires knowledge of at least one infected neighbor. DWM is deterministic in the sense of how many doses are required to become infected. However, DWM is stochastic due to the way it chooses to apply a dose. Therefore, DWM belongs in the same class as ICM in our conceptual framework. Both SLATM and SLFTM are stochastic and their neighbor knowledge requirements retain the same as their deterministic versions. Therefore, SLATM belongs to the same class as ICM, and SLFTM belongs to the same class as BRRM.

Table 3.2: Conceptual framework of models with examples

| | | Stochasticity | |
|---|---|---|---|
| | | Deterministic<br>$G = k$ | Stochastic<br>$G \sim U(0,1)$ |
| Neighbor Knowledge of Infection Requirement | Partial<br>$F = f(A_x \subseteq N_x)$ | Class I<br>LATM | ClassII<br>ICM, DWM, SLATM |
| | Complete<br>$F = f(N_x)$ | Class III<br>LFTM | Class IV<br>BRRM, SLFTM |

This conceptual framework justifies the model selection in this work and ensures that the mechanisms driving the selected models have different underlying conceptual groundings.

### *Experiments*

In order to test the differences between models we focused on analyzing the final states and the speed of simulations as response variables. For analyzing the final states of simulations we mea-

[3]It could also be argued that DWM is a multi-compartmental model that distinguishes from SI due to its property of storing a state variable, namely, the dose of exposure

sured: 1) the final spread of infection, as the fraction of nodes infected, and 2) the final fraction of $I \rightarrow S$ and $S \rightarrow S$ edges, which describes how the transmission occurred. The network is directed.[4] Note that the network may contain four different types of edges according to the state of nodes: an infected node to a susceptible node ($I \rightarrow S$), a susceptible node to another susceptible node ($S \rightarrow S$), susceptible node to an infected node ($S \rightarrow I$), and an infected node to another infected node ($I \rightarrow I$). The $I \rightarrow S$ edges are the most important for propagation of information. The next most important edges for the propagation are the $S \rightarrow S$ edges since they could become $I \rightarrow S$ edges. The other two types of edges ($I \rightarrow I$ and $S \rightarrow I$) do not contribute to further information propagation in the discussed models, since they do not affect the overall diffusion. Therefore only the $I \rightarrow S$ and $S \rightarrow S$ edges are considered in this study.

Since some models (e.g. BRRM and DWM) are designed to run until all reachable nodes are infected, comparison of final state alone may not provide a meaningful comparison of their differences. Therefore, the following measures were implemented to allow us to compare and contrast models. The number of time-steps taken by a model to reach a given amount of network infection has been used in previous research to investigate the speed of information propagation [21, 33, 37]. Another way of looking at speed of propagation is by using Net Present Value (NPV). Stonedahl et al. [46] used NPV as a measure for comparing effectiveness of different seeding strategies for viral marketing, where both speed an quantity of spread was important. Thus, for analyzing the speed of simulations[5] we measured: 1) the number of time-steps taken to reach different stages (such as 50% of infection, and final state) of simulation, and 2) the NPV of infection.

The following experiments were designed mainly to test primary hypotheses that the final state spaces described by each model are different and that the speed of convergence of models are

[4]Therefore an edge going from an infected node to a susceptible node ($I \rightarrow S$) is distinctly different from an edge going from a susceptible node to an infected node ($S \rightarrow I$).

[5]We are grateful to a reviewer who recommended to incorporate the speed of infection as a measure in this study.

different. Our first experiment is a factorial exploration of the parameter space of the models which we denote as EXP1. The parameter value ranges of EXP1 are such that they were varied in the range [0.05, 0.95] with step size of 0.05 except for $p_\omega$ in the range [0.05, 0.95] with step size of 0.45, $k$ in range [2,10] with step size of 1, $\psi$ in range [1, 20] with step size of 1, and $\phi_0$ in range [0.05, 0.35] with step size of 0.05. Each of these configurations was run with 30 replicas. The following measured outcomes were recorded:

1. Final fraction of infection ($\phi_F$) : The number of infected nodes divided by the total number of nodes at the end of each simulation was measured as $\phi_F$.

2. Final fraction of infected to susceptible edges ($I \rightarrow S$) : The number of $I \rightarrow S$ edges divided by the total number of edges

3. Final fraction of susceptible to susceptible edges, ($S \rightarrow S$) : The number of $S \rightarrow S$ edges divided by the total number of edges

4. Number of infections occurred at each time-step.

We also performed a sensitivity analysis to examine how the model parameters affect the final fraction of infection. Our aim in this sensitivity analysis was to investigate whether the variance of model specific parameters of the chosen models possess adequate enough ability to affect simulation outcome with respect to the other common variables(i.e. network parameter and initial infection). Through this analysis we aimed to compare the effects of model parameters of different models against each other. Techniques such as traditional design of experiments approaches, are not adequate for agent-based models due to various reasons such as non-linearity [40]. The overview of analysis methods for agent-based models published by Lee et al. [28] shows that variance based sensitivity analysis as the best approach for investigating the sensitivity of model outcomes to its parameters. Therefore, we used the Sobol sensitivity analysis–a variance based

sensitivity analysis method [45, 39, 30] for identifying the sensitivity of final state to the model parameters. Sobol sensitivity analysis is a method for measuring influence of input variables on the output of a model. It quantifies the contributions of each input variable, both individually and in combination with other variables (i.e. interactions), to the overall variability of the model output. It is based on the concept of variance decomposition, in which the total variance of the model output is partitioned into contributions from individual variables and their interactions. The results of a Sobol sensitivity analysis outputs sensitivity indices such as the first-order sensitivity index, which represents the influence of a single variable (i.e. influence of an individual term), and the total sensitivity index, which captures the influence of a variable and all its interactions with other variables. By these sensitivity indices, we can identify the most influential variables and their relative importance. For each parameter $m$ of a model, this method calculates an index that represents the total sensitivity of the parameter ($S_{T,m}$) which is a representation of the total contribution of the parameter to the variance of the response variable. Further, this method provides the fraction of contribution by the $1^{st}$ order term of model parameter ($S_{1,m}$), and the fraction of contribution by $2^{nd}$ order interaction terms of individual model parameters ($S_{2,m}$) to the variance of the response variable.

Following the method employed by Ligmann-Zielinska et al. [30], we designed Sobol sensitivity analysis experiment using low-discrepancy Saltelli sampling of parameters with 4096 samples per parameter (i.e. more than 16000 simulations per model). Using this design we conducted two experiments analyzing sensitivity of $\phi_F$ and NPV to model parameters of each model. We denote these two experiments as EXP2 and specifically we denote the experiment of analyzing sensitivity of $\phi_F$ and NPV as EXP2a and EXP2b respectively.

*Simulations:*

All models were implemented in Python programming language. The number of nodes was set to 5000 for all experiments. Each simulation was run for a maximum of 10000 time-steps with 30 replicas per each parameter configuration. At each $1000^{\text{th}}$ time-step (e.g., 1000, 2000, 3000, etc.) we check if there was at least one node that changed its state during the last 1000 time-steps, and if there was no state changes in the last 1000 time-steps then we stop the simulation. The outputs included the fraction of infected nodes, the fractions of $I \rightarrow S$ edges, $S \rightarrow S$ edges, and number of infections at each time-step. The python library SALib [23] was used to conduct the Sobol sensitivity analyses in the EXP2. The boundaries of parameters for Saltelli sampling for the EXP2 was matched to the parameter ranges used in EXP1.

*Network Generation:*

As mentioned in the literature reveiw, in the network science literature there are many different network structures that have been studied, such as random networks [19], small-world networks [52], and scale-free networks [4]. The dynamics of the diffusion of information on these various types of networks has been the focus of a number of studies [51, 14, 26]. Since scale-free networks (SF) are well studied and comparable to real world networks, we chose to investigate the information diffusion over SF networks in this study. Moreover, since the information flow between nodes could be asymmetric, i.e., many users may follow celebrities on social media, but celebrities follow very few users, we specifically use directed scale-free networks for this study. The original scale-free network generation algorithm proposed by Barabási et al. [4] generates undirected networks, so instead we use the algorithm proposed by Bollobás et al. [7] for generating directed scale-free networks for our experiments. A directed network can contain unreachable nodes (nodes without incoming edges) and terminal nodes (nodes without outgoing edges). Since such nodes are also

a part of the system, they were not removed. In our models, terminal nodes can change state, but can not influence the state of others, while unreachable nodes can not change state, but could affect the state of others. The Bollobás et al. algorithm has three parameters: $\alpha$, $\beta$ and $\gamma$ which are probabilities such that $\alpha + \beta + \gamma = 1$. The parameter $\beta$ is the probability that a new edge is added to the network between two existing nodes while the other two parameters $\alpha$ and $\gamma$ define the probability of adding a new node to the network through an outgoing-edge and incoming-edge respectively. The three probability values are used to control the density and connectivity of the generated network. Therefore, $\beta$ governs the edge density of the network and also affects connectivity (increasing $\beta$ increases the edge density). In this study, we observed the dynamics of various DOI models over different $\beta$ values for a fixed number of nodes while keeping $\alpha$ and $\gamma$ equal such that $\alpha = \gamma = \frac{1-\beta}{2}$.

## Results

Initially, we examine the outcomes from EXP1, which detail the distributions of nodes and edges in the DOI models. Subsequently, we explore the results of the sensitivity analysis from EXP2.

### *Node and edge distributions at the final state and speed of DOI*

This subsection describes the results obtained from the EXP1 experiment. The violin plots[6] of the final fraction of infection ($\phi_F$) are given in Fig. 3.1. The normality test for the final fraction of infection using Anderson-–Darling tests confirmed that none of these $\phi_F$ distributions are normal. We compared the $\phi_F$ distributions of models against each other using Wilcoxon rank-sum tests. It is a non-parameteric statistical test that compares two independent samples. It assesses whether the

[6]These are violin plots with embedded box plots which shows kernel density and spread. The kernel density shape depicts information of peaks while box plot captures locality, spread, and skewness of data

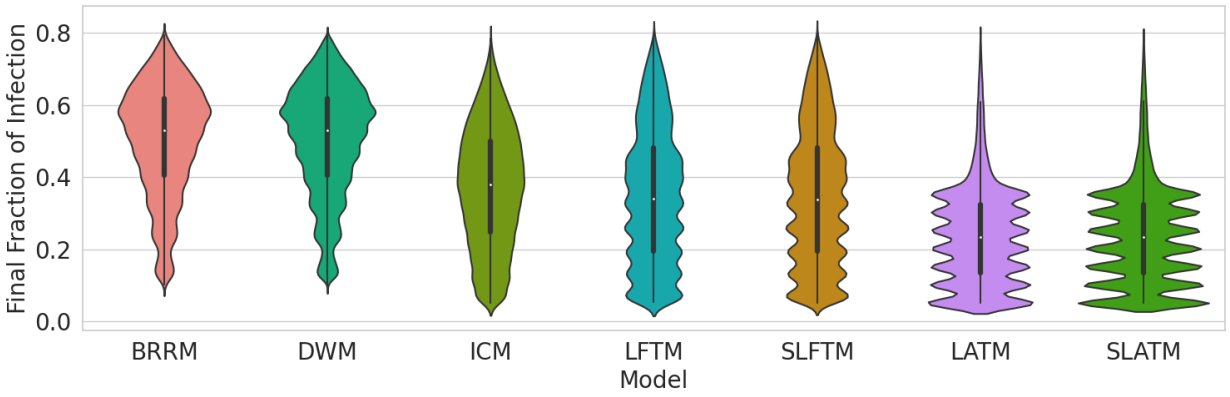


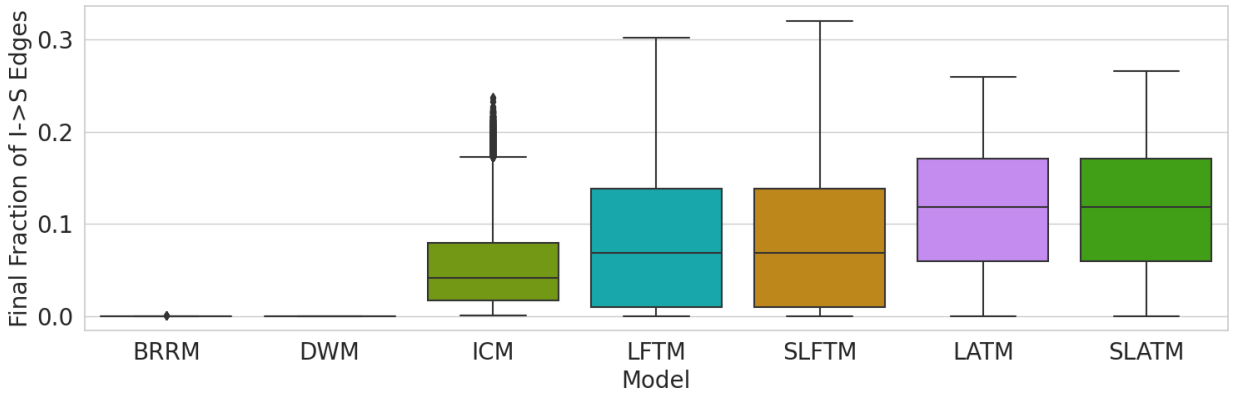

Figure 3.1: Distribution of final fraction of infection of models

Table 3.3: Mean, Std. dev., Skewness and Kurtosis of Distributions of the Final Fraction of infection of Models

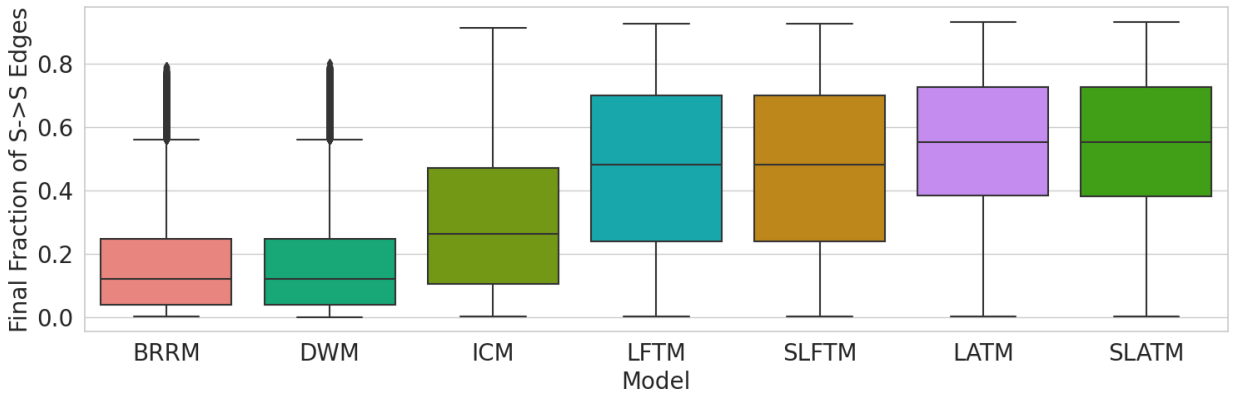
| Model | Mean | Std. Dev. | Skewness | Kurtosis | Excess Kurtosis |
|---|---|---|---|---|---|
| **BRRM** | 0.503639 | 0.152363 | -0.580979 | 2.661265 | -0.338735 |
| **DWM** | 0.503608 | 0.152343 | -0.580534 | 2.659739 | -0.340261 |
| **ICM** | 0.373487 | 0.164751 | -0.038233 | 2.177251 | -0.822749 |
| **LFTM** | 0.350678 | 0.185220 | 0.223268 | 2.092012 | -0.907988 |
| **SLFTM** | 0.350518 | 0.185232 | 0.224638 | 2.092606 | -0.907394 |
| **LATM** | 0.236928 | 0.134648 | 0.688209 | 3.540513 | 0.540513 |
| **SLATM** | 0.236932 | 0.134686 | 0.689483 | 3.543973 | 0.543973 |

medians of the two groups are significantly different. This test is appropriate when the data do not follow a normal distribution. For comparisons of BRRM vs DWM, SLFTM vs LFTM, and SLATM vs LATM we received very high p-values (0.95, 0.84, 0.99 respectively), and for all else p-values were zero. Therefore, those mentioned pairs of models have statistically similar $\phi_F$ distributions. The mean, standard deviance, skewness, and kurtosis of each $\phi_F$ distribution of the models are given in Table 3.3. Inspection of both the Fig. 3.1 and Table 3.3 ensures that BRRM and DWM have left skewed $\phi_F$ distributions, and LATM and SLATM have right skewed $\phi_F$ distributions.

Notice that the shape of $\phi_F$ of LATM, SLATM, LFTM, and SLFTM have some jaggedness (Sine wave like patterns). This is likely due to the discreteness of the infection condition. The infection




(a) Distributions of fraction of $I \rightarrow S$ edges at final state




(b) Distributions of fraction of $S \rightarrow S$ edges at final state

Figure 3.2: Distributions of edge types at final state

condition of these models are directly related to number of infected neighbors, which is a discrete value regardless of whether the threshold is a fraction or a whole number.

At the end of each simulation the fraction of each edge type was calculated. Fig. 3.2 shows these distributions. The distributions of these edge types are results of the simulation parameters, the model, and the network. Together they describe the infectiousness of the network at the end state of all simulations. The fraction of $I \rightarrow S$ edges at the final state represents the number of edges that failed in the propagation (hence it still remains as $I \rightarrow S$). Similarly, $S \rightarrow S$ edges in the final state represents edges that could have been used to propagate information, but were not.

We compared the distributions of both of these edge types between all model pairs using Wilcoxon rank–sum tests and found that, BRRM vs DWM, LFTM vs SLFTM, and LATM vs SLATM have very high p-values and for all other pairs the p-values were zero. For values near zero, the models produce different final states, but for the models with high p-values, the end results are not easily differentiated. Therefore, similarly to the $\phi_F$ distributions, we conclude that distributions of $I \to S$ and $S \to S$ edges resulted from the pairs BRRM vs DWM, LFTM vs SLFTM, and LATM vs SLATM are similar.

Visual inspection of the Fig. 3.2 also shows that BRRM and DWM does not contain $I \to S$ edges and they have the lowest amount of $S \to S$ edges left at the final state. This is due to the fact that BRRM and DWM runs until all reachable nodes are infected, thus the behaviour of these two models are different from the other models.

By considering the differences between both the final state of edge distributions and the final state of spread ($\phi_F$) we can conclude that the final state space of ICM, LATM, LFTM, and BRRM are significantly different from each other. Moreover, we conclude that final state space of the pairs BRRM & DWM, LFTM & SLFTM, and LATM & SLATM are similar.

Since, BRRM and DWM run until all reachable nodes are infected, when these models have finished a run, the set of infected nodes is equivalent to the set of nodes that are reachable from the initially infected nodes, regardless of the DOI model's parameters. Therefore, $\phi_F$ distributions of BRRM and DWM are actually representing the fraction of nodes in the reachable network. As a result, the $\phi_F$ distributions observed after running DWM and BRRM are not determined by the parameters of the DOI model, instead they are determined by the network generation algorithm. Similarly the $I \to S$ and $S \to S$ edge distributions of BRRM and DWM are also determined by the network generation algorithm. Therefore, it is not possible to have a meaningful comparison between BRRM and DWM by just looking at the final state of nodes and edges. So, to create

a meaningful comparison between models such as BRRM and DWM, we look at the "speed" of infection.

Table 3.4: Mean and maximum number of time-steps taken for models to reach different stages of simulations

| | $\phi_F \geq 50\%$ by | | $\phi_F \geq 65\%$ by | | $\phi_F \geq 75\%$ by | | Halt by | |
|---|---|---|---|---|---|---|---|---|
| Model | mean | max | mean | max | mean | max | mean | max |
| LATM | 1.71 | 8.0 | 2.28 | 6.0 | 2.61 | 5.0 | 2.77 | 18 |
| ICM | 1.71 | 12.0 | 2.26 | 6.0 | 2.62 | 5.0 | 5.73 | 19 |
| LFTM | 2.29 | 43.0 | 3.25 | 41.0 | 3.21 | 11.0 | 6.12 | 52 |
| SLATM | 15.96 | 201.0 | 23.98 | 254.0 | 31.89 | 178.0 | 25.41 | 502 |
| SLFTM | 25.98 | 986.0 | 40.75 | 797.0 | 49.60 | 307.0 | 91.06 | 1184 |
| BRRM | 11.69 | 961.0 | 14.90 | 884.0 | 17.64 | 671.0 | 126.85 | 3004 |
| DWM | 37.25 | 2054.0 | 50.41 | 1577.0 | 60.27 | 951.0 | 134.81 | 3093 |

Sorted by the mean number of time-steps taken to halt the simulation.

We measure the number of time-steps taken to infect 50%, 65% and 75% of nodes, the number of time-steps taken to halt the simulation, and the NPV of the infections. We define the number of time-steps taken to halt the simulation as the number of time-steps taken to reach the final state of the simulation[7]. The Table 3.4 shows the number of time-steps taken by each model to reach the different stages of simulations as described above[8]. LATM, ICM and LFTM are the fastest models at all stages compared to the other four models. DWM is the slowest model to reach every stage shown in the table. BRRM is the next slowest model to reach the final state. However, BRRM is only slower when comparing the amount of time taken to reach end of simulation. BRRM reaches 50%, 65%, and 75% of nodes sooner than both SLATM and SLFTM on average case. On the contrary, the max number of time-steps to reach 65%, and 75% of nodes show evidence that BRRM might take longer times than SLATM and SLFTM to reach those stages. Another observation from

[7]The simulation might run past this time-step in order to identify that there was no change in the system as described in Section 3

[8]Note that not all simulations will reach 75% infection (or even 50% or 65%) due to their parameter values being less contagious (e.g., an ICM run with very low $q$ and $\beta$ values might not reach 75% infection). Such simulation runs were omitted when calculating the mean time-step values of that respective percentile.

this table is that both linear threshold models (LATM and LFTM) are significantly faster at reaching every stage than their respective stochastic versions (SLATM and SLFTM). Therefore it is evident that different models have different speeds of information spread at different stages of simulations. In other words, while one model may be slower than another model in reaching a particular stage, the same model may demonstrate faster progress than the other model in reaching a different stage.

NPV is able to give a summary value which represents both the quantity (i.e., number of infected nodes) and the speed (i.e., number of time-steps spent) of infection. The basic idea is that an adoption of the infromation today is worth more than adoption tomorrow. This is an important aspect especially for information campaigns. For example, in a disinformation campaign, the speed of information propagation is important since one party wants to propagate some information before the opposition spreads a counter argument. Therefore, the amount of spread that can be achieved "today" is more valuable than the amount of spread that we could have by "tomorrow". We calculate NPV (with deprecation factor $\lambda = 90\%$ and profit factor $p_f = 1$) using the Eq. 3.2 which is adopted from Stonedahl et al. [46]. $t$ is the time-step index using zero based indexing. $a_t$ is the number of infections at time-step $t$.

$$NPV = \sum_{t=0}^{\infty} a_t p_f \lambda^t = \sum_{t=0}^{\infty} a_t (0.9)^t \tag{3.2}$$

The NPV distributions for each model are shown in Fig. 3.3. The NPV distributions of BRRM, ICM, and LFTM are larger than the other models. Therefore, these results help in distinguishing those three models from the rest of the models. Combined with the results shown in Table 3.4 we identify that BRRM infects more nodes than DWM in the early time-steps (The NPV of BRRM is higher then DWM and BRRM is faster at reaching all stages in Table 3.4 than DWM.).

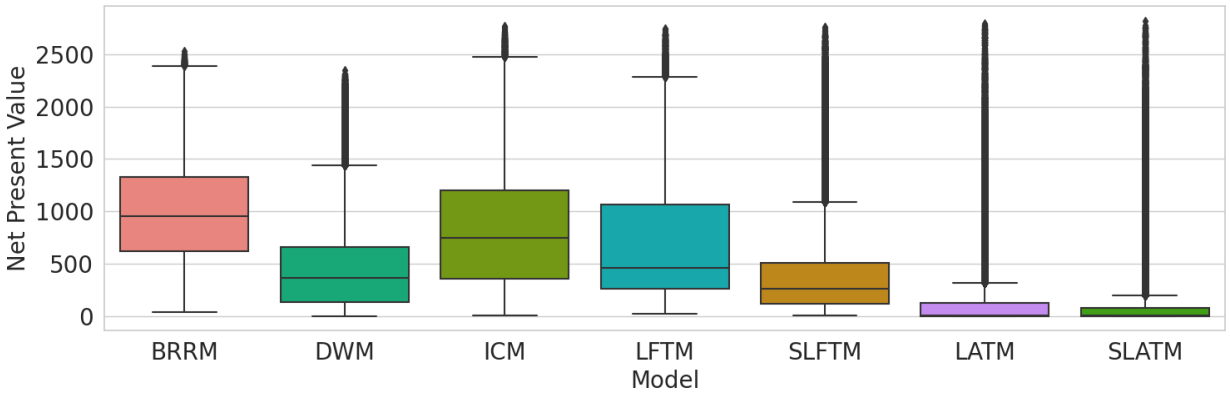
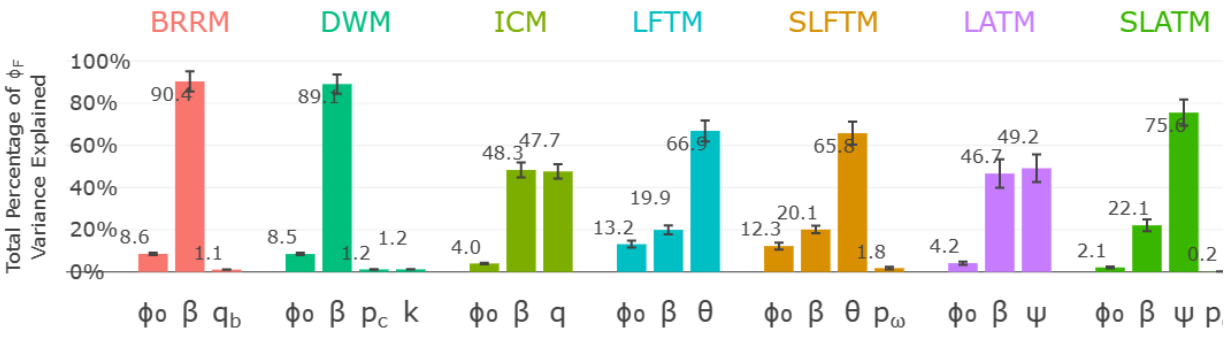


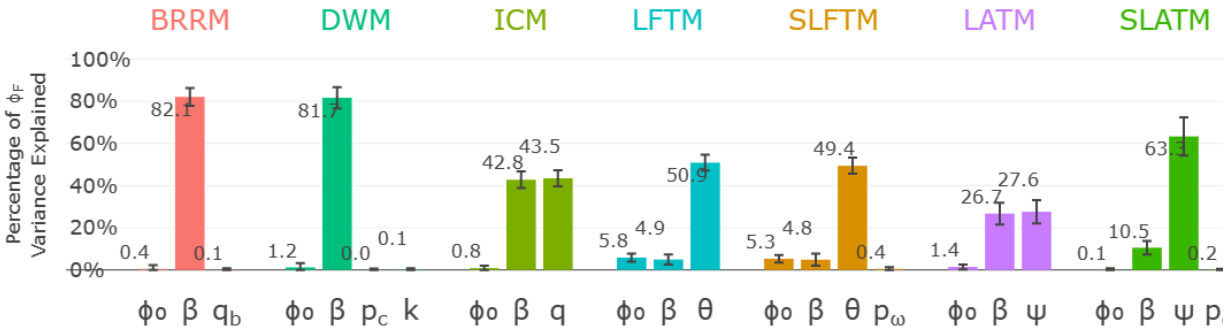

Figure 3.3: Distributions of net present value of model runs

*Sensitivity analysis of the spread and speed of diffusion*

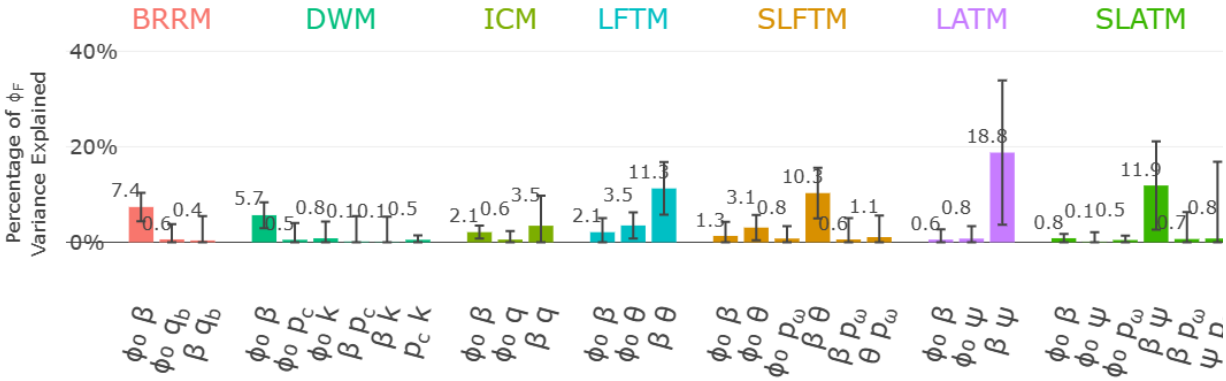
This subsection describes the results obtained from EXP2 experiments. In EXP2a, we used Sobol sensitivity analysis to observe the effects of model parameters on the variance of $\phi_F$. The decomposition of $\phi_F$ variance for the models are given in Fig. 3.4. Our results of the Sobol sensitivity analysis contains three parts: total sensitivity index ($S_T$), 1$^{st}$ order sensitivity index (also called main effect index) ($S_1$), and 2$^{nd}$ order sensitivity index ($S_2$)[9]. The $S_1$ is a measure of individual model parameters and $S_2$ is a measure of the interaction of two parameters. The $S_1$ of a given model parameter depicts the fraction of contribution given by independent term of the parameter to the variance of the outcome. Similarly $S_2$ of a given pair of model parameters depicts the fraction of contribution given by the interaction of those two model parameters to the variance of the model outcome. The $S_T$ of a model parameter depicts the overall contribution of the parameter on the outcome variance.




(a) Confidence Intervals of $S_T$ Decomposition of Variance of $\phi_F$




(b) Confidence Intervals of $S_1$ Decomposition of Variance of $\phi_F$




(c) Confidence Intervals of $S_2$ Decomposition of Variance of $\phi_F$

Figure 3.4: Sensitivity of Final Fraction of Infection ($\phi_F$)

*Sensitivity analysis based on $\phi_F$*

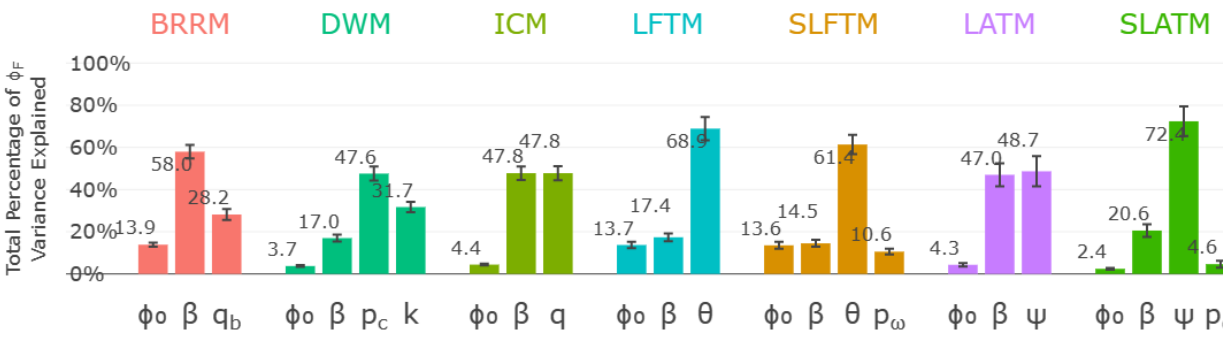
The $\phi_F$ of BRRM and DWM have similar sensitivity to parameters such that the final fraction of infection is about 90% dependent on the network parameter $\beta$ and a 8.5% dependent on the initial

[9]We are adopting the abbreviation presented in [44, 30]

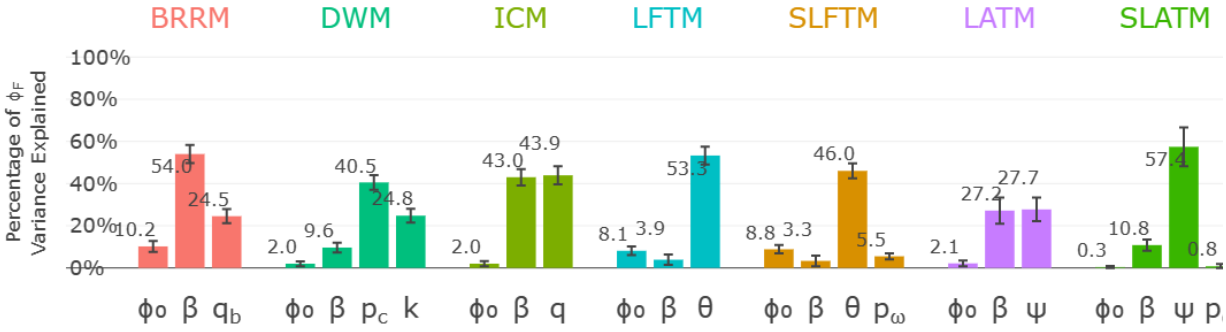
fraction of infection $\phi_0$ (Fig. 3.5a). As further decomposition given in Fig. 3.5b and Fig. 3.5c shows, more than 85% of their $\phi_F$ variance is determined by the first order term $\beta$ ($\gtrapprox$ 81% in Fig. 3.5b) and the second order term $\phi_0\beta$ ($\gtrapprox$ 5% in Fig. 3.5c). This proves that the $\phi_F$ of these two models are almost completely governed by the network parameter and the initial infection. This is due to BRRM and DWM being run until all reachable network is infected. The $\phi_F$ of ICM shows equivalent total sensitivity to both the model parameter $q$ and network parameter $\beta$ ($\approx 48\%$ each) and shows low sensitivity (about 4%) to $\phi_0$ (Fig. 3.5a). Both LFTM and SLFTM models show high sensitivity ($\approx 66\%$) to their model parameter $\theta$. SLFTM shows only tiny sensitivity to its parameter $p_\omega$ which seems almost negligible. Therefore, the sensitivity of model parameters to the $\phi_F$ in LFTM and SLFTM are similar. Interestingly the LATM and SLATM models do not show similarities, although their final states which we observed in the EXP1 were similar. The LATM model shows approximately similar sensitivities to network parameter $\beta$ and it's model parameter $\psi$. The second order parameter $\beta\psi$, which combines those two parameters, is showing a considerably large effectiveness in determining the outcome of LATM as well. Therefore, we could say that both $\beta$ and $\psi$ are equivalently effective in determining the outcome of LATM model. SLATM model shows highest sensitivity to it's threshold $\psi$ and secondly to $\beta$, as evident from all three Fig. 3.4. SLATM doesn't seem to be sensitive to it's other model parameter $p_\omega$.

*Sensitivity analysis based on NPV*

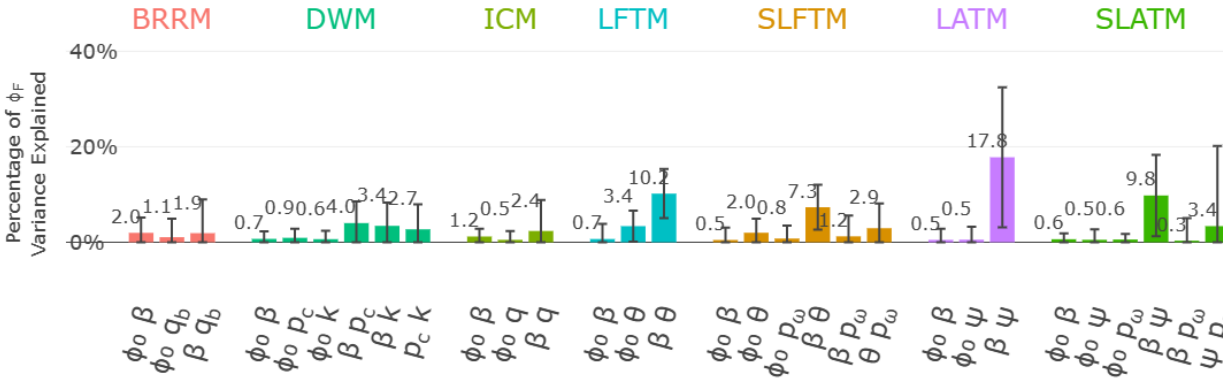
In EXP2b, we evaluated the sensitivity of parameters to the NPV of infection. The results of these experiments are shown in Fig. 3.5. There are two important things to notice about the Fig. 3.5 : (1) Results have approximately similar sensitivity values as Fig. 3.4 for all models except BRRM and DWM. (2) The sensitivity decomposition of BRRM and DWM have unique values. The sensitivity index values of ICM, LFTM, SLFTM, LATM, and SLATM are almost exactly similar to their respective values from EXP2a, which indicates the consistency of those models in responding




(a) Confidence Intervals of $S_T$ Decomposition of Variance of NPV




(b) Confidence Intervals of $S_1$ Decomposition of Variance of NPV




(c) Confidence Intervals of $S_2$ Decomposition of Variance of NPV

Figure 3.5: Sensitivity of Net Present Value (NPV)

to the parameters with respect to both speed and spread of infection. The sensitivity values of BRRM and DWM are very different from their respective values in EXP2a. A 58% of variance of the NPV of BRRM is governed by $\beta$. Its model parameter $q_b$ exhibits only half ($\approx$ 28%) the effectiveness of $\beta$. For DWM, more than 78% of NPV variance is dependent on model parameters

$p_c$ and $k$. Therefore, the sensitivity decomposition of BRRM and DWM with respect to NPV are completely different. This difference between sensitivities of BRRM and DWM model parameters was expected since from EXP1 we found that BRRM converges faster than DWM.

## Discussion

Through these experiments we have compared the seven models in a variety of ways. In this section we are specifically going to highlight some of the most interesting differences and similarities.

Let us first look at why DWM seem to produce similar results as BRRM while yet being very different from each other. As discussed, both models run until all reachable nodes becomes infected. Therefore, differences we could observe lie on the speed of information propagation and in mid states of the simulation. Table 3.4 shows that DWM is the slowest model of among these models. This is understandable since DWM has a long process for generating an infection compared to other models because each interaction occurs with a probability of contact ($p_c$) and a contact only adds some dose of exposure to the node. A $k$ number of such doses are required for the node to become infected. Therefore, a susceptible node with any number of infected neighbors requires a minimum of $k$ time-steps to become infected in the DWM. In contrast to the DWM, all other models discussed in this work allow a susceptible node to become infected within a single time-step if a sufficient number of its neighboring nodes are infected. Observing both the Table 3.4 and Fig. 3.3 we summarize that BRRM infects a greater number of nodes within a given number of time-steps as compared to DWM. The parameters of BRRM and DWM are only able to control the speed of the information propagation. This result could be generalized such that, the model specific parameters of a DOI model, which infects all reachable nodes, are only useful in changing the speed of infection.

Let us look at the stochastic versions of LATM and LFTM models: SLATM and SLFTM. Results in EXP1 show that both of these models produce $\phi_F$, and $I \rightarrow S$ and and $S \rightarrow S$ edge distributions which are equal to the results produced by their deterministic counter parts (Fig. 3.1 and Fig. 3.2). However in comparison to their deterministic counterparts, stochastic models take significantly large number of time-steps to converge to each tested simulation stage (Table 3.4). This is an expected behavior since the only difference of these two stochastic models from their respective deterministic counterparts are the probability based conditional checking that was added as a final step of the decision rule. Since this final step is executed at every time-step, eventually any node that was blocked from becoming infected solely due to this final step will become infected. Therefore, if an LATM and SLATM was run on two identical networks with the same initial infected nodes and same model parameter $\psi$ values, the final states of the network nodes and edges will be identically same regardless of the $p_\omega$ value. Similarly this is true for LFTM and its stochastic version: SLFTM. The sensitivity analysis results also have confirmed that the stochastic model parameter $p_\omega$ has no control over the outcome $\phi_F$.

Interestingly for all models except for BRRM and DWM, we observed that the sensitivity of model parameters with respect to NPV have the same values as sensitivity against $\phi_F$. We speculate that a correlation between the speed and spread of diffusion might be the reason for this behaviour. We conclude that further investigation is required for uncovering the reason behind this phenomenon, which could potentially be a future work. However, the most focused result in here was that the model parameters of BRRM and DWM were able to exhibit their unique characteristics (without being obscured by the fact that they run until all reachable network is infected) when sensitivity was measured against NPV. Due to this reason NPV could be identified as a superior measurement when it comes to sensitivity analysis of DOI models.

When considering all figures and results from EXP1, we see that LATM struggles in producing

a large number of infections[10]. A reason could be because the condition for LATM to produce infections expects a minimum number of infected neighbors irrespective of the neighborhood size. The mean degree overall for the conducted simulations is 5. So it could be that some threshold parameters such as $\psi \geq 6$ are too high for those networks with average in degree of 5. Notice that the range of values used for $\psi$ was [1, 20].[11] This is the reason why there are many underutilized $I \rightarrow S$ edges left in the LATM model as seen in Fig. 3.3a.

The ICM shows more of a balanced distribution of $\phi_F$ compared to all other models (Fig. 3.1 and Table 3.3). The $I \rightarrow S$ and $S \rightarrow S$ edge distributions of ICM are lowest when compared to all except DWM and BRRM. Therefore, apart from the models that run until all reachable network becomes infected, ICM is the model that could utilize most of the edges in the network and therefore has the potential to infect most of the network (Fig. 3.2). When comparing speed of diffusion, ICM is at the 2$^{\text{nd}}$ fastest, losing only marginally to LATM (Table 3.4). ICM has a NPV distribution that is larger than all except BRRM and LFTM (Fig. 3.3). The conducted sensitivity analysis shows that ICM performs similarly against both $\phi_F$ and NPV. The model parameter $q$ and network parameter $\beta$ have equal control over the outcome of ICM. Therefore, we conclude that ICM has the ability to produce models for a wide variety of final infection and speed of infection states. Therefore, ICM could be considered as a general purpose model. An additional advantage would be that since ICM runs in a small number of time-steps, running experiments with ICM is computationally efficient. Since there is only one parameter that exists that is specific to the ICM model, a possible limitation of ICM would be that changing the model parameter will change both the speed and final state in tandem.

The LFTM and SLFTM show comparable results of $\phi_F$ distributions compared to ICM. When

[10] So does SLATM since these issues of LATM related to its final state will shadow the final state of SLATM.

[11] Our findings indicate that the conclusions on $I \rightarrow S$ and $S \rightarrow S$ edge distributions remain valid even when the value of $\psi$ is less than 6.

considering edge distributions, LFTM and SLFTM are similar to ICM in their capacity to utilizing edges. LFTM has high NPV and requires lesser time-steps to run. SLFTM has lower NPV and takes comparatively large number of time-steps to run. While performance of LFTM seems to follow ICM, the sensitivity index of LFTM parameters exhibit to be very different. Most of the LTFM outcomes ($\approx 68\%$) are determined by $\theta$, and the network parameter $\beta$ only has less than 20% of control over the outcomes. While the $\theta$ parameter of LFTM affects both the rate of spread and the final state, SLFTM has the ability to affect the rate of spread independently by varying $p_\omega$ [12].

Overall, we have found that using our generalized framework we can classify how concepts for different models could be classified into a common framework. However, when we investigate the final state, speed of infection propagation, and sensitivity to final fraction of spread, we found that these models behave in unique and different ways regardless of the class it belongs to in the conceptual framework. Even though two models might belong to the same conceptual class they can produce completely different outcomes from their simulations (e.g. BRRM vs SLFTM, ICM vs DWM, and ICM vs SLATM). Moreover, we have found that regardless of being conceptually different (belonging to two different conceptual classes), two models may produce similar simulation outcomes (e.g. BRRM vs DWM, LFTM vs SLFTM, and LATM vs SLATM).

[12] Similarly the rate of spread of SLATM is independently affected by $p_\omega$

# CHAPTER 4: QUANTIFYING THE EFFECT OF CLUSTERING COEFFICIENT ON MODELS OF INFORMATION DIFFUSION

In this chapter we describe the second study which is a direct extention of the previous study. This study aims to investigate the influence of network node clustering on the dissemination of information.

## Methodology

Network clustering is quantified using the Clustering Coefficient (CC) as introduced by Watts and Strogatz[52]. Information propagation is typically determined by the fraction of infected nodes, denoted as $\phi_F$. However, given the significant role of time in the spreading process, we selected Net Present Value (NPV) as the metric to quantify the spread. NPV is calculated by using the equation 3.2. We applied a depreciation factor of $\lambda = 90\%$ and a profit factor of $p_f = 1$, following the methodology outlined by Stonedahl, Rand, and Wilensky[46].

The complex contagion model described by Centola, Eguíluz, and Macy[10] is a threshold based model similar to LFTM. The simple contagion model described by Watts and Strogatz[52] is a probability based model similar to ICM. As there are multiple types of DOI models available in the literature, we investigate three specific classical models of information diffusion: ICM, LATM, and LFTM, to span the full range in our simulations. In the ICM, each infected node has a single opportunity to infect each of its susceptible neighbors, and the probability of infection is determined by the characteristics of the connecting edge. In the LATM, each susceptible node evaluates its threshold concerning the number of infected neighbors in order to determine infection. In the LFTM, a susceptible node becomes infected when the fraction of infected neighbors surpasses its

threshold value.

The aim of this research is to examine the following hypotheses:

1. The CC's effect on NPV/Final Infection depends on the Model Types
2. The CC's effect on NPV/Final Infection depends on the Model Parameter
3. The CC's effect on NPV/Final Infection depends on the Network Parameter

### *Experiments*

Our experimental design focused on generating results from three categories of networks and three agent-based information diffusion models, each comprising 1000 nodes. These networks were constructed using three distinct generation algorithms: (1) Watts and Strogatz (WS) [52], (2) Erdos and Rényi (ER) [19], and (3) Barabási and Albert (BA) [4]. The probability parameter of Erdos and Rényi algorithm and Watts and Strogatz algorithm were varied in the range $[0.1, 0.45]$ with 0.05 step size. The mean degree $k$ of Watts and Strogatz network generator was kept constant as $k = 5$. The growth parameter $m$ of Barabási and Albert was varied in the range $[4, 7]$ with step size of 1. The both imitation probability $p$ of ICM and the fractional threshold $\theta$ of LFTM were varied in the inclusive range $[0.05, 0.95]$ with 0.05 step size. The threshold $\psi$ of LATM was varied in the range $[1, 10]$ with step size of 1. The initial infection $\phi_0$ was varied in the range $[0.025, 0.35]$ with 0.025 as step size. Each configuration was run with 30 replicas.

### *Implementation*

All simulations were implemented in Python using the `NetworkX` library for network generation algorithms. Each simulation ran for a minimum of 1000 time steps, and the simulation stopped

when there was no change in the network for 1000 consecutive time steps.

*Statistical Analysis*

We are interested in quantifying the relationships between the CC and the NPV. Specifically, we aim to evaluate the average changes in NPV associated with a unit increase in the CC across three model types: ICM, LATM, and LFTM. Additionally, our goal is to compare the variability of NPV as the CC varies between 0 and 0.45 as shown on the left side of Figure 4.1.

The distribution of CC is predominantly influenced by the network type, due to the design of the data generating process. In particular, CC in a SF network is significantly affected by the network parameter, given that clustering is contingent on the probability inherent to the generation process. In contrast, CC in a R network and SW network indicates linear associations with the network parameter. Specifically, it shows a positive correlation for R network and a negative one for SW network, as illustrated on the right side of Figure 4.1.

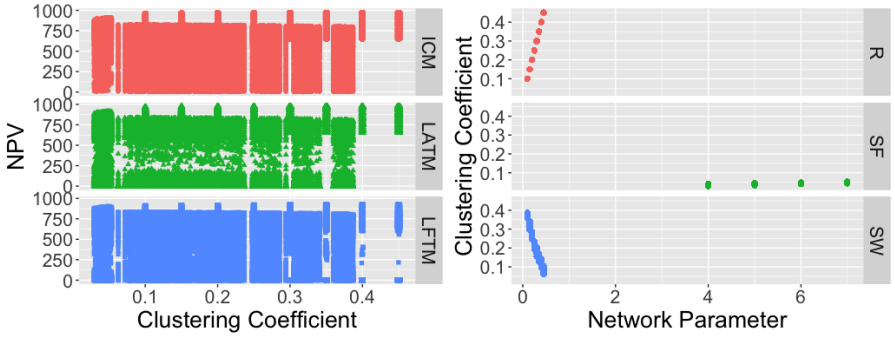



Figure 4.1: Left: CC and NPV by Model, Right: Network Parameter and CC by Network Type

We observe that NPV has a zero-inflated, left-skewed distribution, which is affected by the model parameter. This suggests that a simple parametric distribution may not be suitable for modeling such outcomes [31]. To determine the optimal modeling approach aligning with the study's objectives, we investigated the sources and essential factors contributing to the zero inflation and

left-skewed characteristics of NPV. These characteristics are associated with both the model type and network type.

As illustrated in Figure 4.2 on the left, approximately 20% of all NPV observations have zero NPV values. Among these zero NPVs, around 23% originate from the LATM model, while the remaining approximately 77% are attributed to the LFTM model. This distribution aligns with expectations, given that the ICM model has negligible initial infections, while the LFTM model has the highest initial infection rate across all ranges. The LATM model falls between these extremes, as shown on the right side of Figure 4.2. Furthermore, the number of initial infections gradually decreases as the initial infection rate increases from 0.025 to 0.35. Moreover, the LFTM model consistently exhibits nearly four times as many initial infections at each step compared to the LATM model.

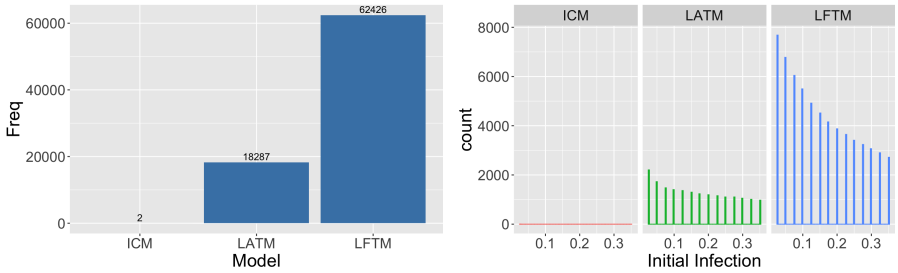



Figure 4.2: Left: Count of Zero NPV by Model Types, Right: Initial Infection by Model Types for Zero NPV

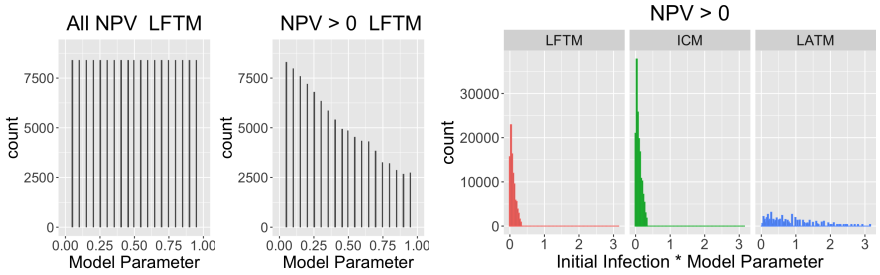



Figure 4.3: Left: Model Parameter Before and After the Filtering, Right: Initial Infection Times Model Parameter for NPV>0

The left side of Figure 4.3 indicates that a significant proportion of zero NPVs in the LFTM model are associated with larger model parameters, accounting for approximately 39% of all observations within the LFTM model. Similarly, zero NPVs in the LATM model are generated by relatively large model parameters, particularly those toward the end of the interval ranging from 5 to 8. In the LATM model, zeros make up approximately 24% of all LATM model observations, whereas in the ICM model, this percentage is less than 0.01%.

On the right side of Figure 4.3, the distribution of initial infection times model parameters appears similar for the ICM and LFTM models, both following a right-skewed distribution with more values clustered around zero. In contrast, the initial infection times model parameters in the LATM model are more evenly distributed across the range, with a higher frequency of values between 0 and 2, and a lower frequency for values beyond 2.

The sensitivity analysis of NPV and Final Infection related to initial infection, density parameter, and model parameter has been previously studied by Jayalath et al. [25]. Their findings informed the selection of meaningful predictors for our model in this work. According to their research, for the ICM and LATM models, the density parameter and the model parameter each contribute equally to nearly 50% of the outcome, while for the LFTM model, the model parameter accounts for almost 70%, followed by the density parameter at 17%, and the initial infection at around 13%. These proportions appear to hold relatively consistently for both NPV and Final Infection.

As shown in Figure 4.4 on the top left, there is a observable difference in mean NPV between the ICM/LATM models and the LFTM model. However, when considering Non-Zero NPVs (bottom right in Figure 4.4), the average NPV for the LFTM model is not significantly lower compared to the other two models.

Similar trends are observed when examining the impact of network types on NPV. On the top right in Figure 4.4, the mean NPV exhibits substantial variation between the SW network type and the

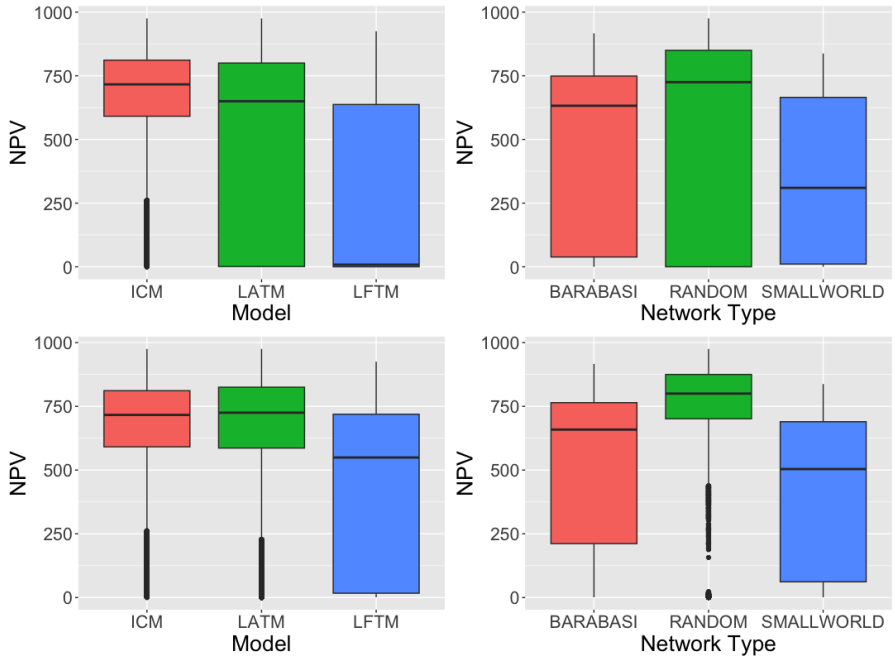



Figure 4.4: Top: All NPV, Bottom: NPV>0

other two network types, but this difference diminishes by approximately half for Non-Zero NPVs (bottom right in Figure 4.4).

In summary, both model and network types prove valuable in predicting zero NPVs within our framework. There are considerable mean differences observed between the LFTM model and the other two models, as well as a similar mean shift between the SW network type and the other two network types. These findings have led us to redefine the reference groups when using the factor predictors Model Type and Network Type. We have selected LFTM and SW as baselines for comparison in Model Type and Network Type, respectively.

Since zero NPV implies no diffusion in the network even after 1000 time-steps in our case, and the focus of this study is to evaluate the associations between CC and NPV when some level of diffusion occurs, we primarily concentrate on Non-Zero NPV for this work. Even after filtering out observations with zero NPVs, the distribution of Non-Zero NPV remains left-skewed, with a

substantial number of 1s on the lower end. These 1s constitute approximately 1.8% of the data and will be included in the analysis, as they indicate some level of diffusion.

To interpret the associations between CC and NPV, we apply Generalized Additive Modeling (GAM) [54] where non-linear relationships are allowed and can be captured through a sum of functions of each feature. In this modeling approach, the response variable is Non-Zero NPVs, and the predictor variable is mean CC adjusted with Model Type, Network Parameter, Model Parameter, and the interactions between mean CC and the two parameters. We have selected the LFTM model and SW network type as the reference levels because we anticipate more substantial mean differences between the LFTM model and the other two models regarding Non-Zero NPVs. Additionally, we expect an evident contrast between Network Type SW and the other two Network Types.

In order to find a good fit for the existing response and preserve interpretability in line with the goal of the study, we use the original scale of NPV as well as the mean CC. We fit a GAM, which can accommodate the non-linear relationships between the predictors and the response[54], as follows:

$$\begin{aligned} NPV_i = &\alpha + f_1(CC\_i) + f_2(Model_i) + f_3(NetParameter_i) + \\ &f_4(ModelParameter_i) + f_5(CC\_i, NetParameter_i) + \\ &f_6(CC\_i, ModelParameter_i) + \varepsilon_i \end{aligned} \tag{4.1}$$

where $\alpha$ is the intercept parameter, the $f_j$ (for $j = 1, \cdots, 6$) are smooth functions, and the $\varepsilon_i$ are independent $N(0, \sigma^2)$ random variables. In our case, as the model is a categorical variable with three levels, it will lead to a parametric estimate indicating the average differences compared to the reference model LFTM regarding their contribution to Non-Zero NPVs. We include the model parameter and network parameter, as both are related to the NPV values during the data generating process. All smooth terms are fitted using cubic splines with 10 basis functions along with a second

derivative-based penalty. We denote this as Model 1, where the response is Non-Zero NPV, and the predictors are listed below.

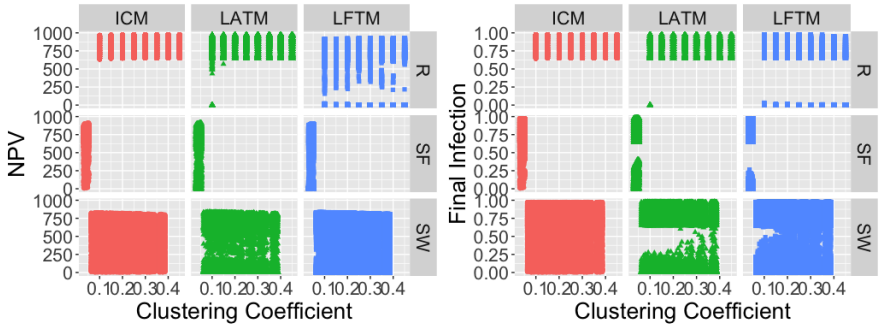



Figure 4.5: CC and NPV/Final Infection by Model and Network Type

Additionally, we applied the same model to the Final Infection, as initial observations from Figure 4.5 suggest that the CC has a more pronounced impact on Final Infection in relation to its effect on Non-Zero NPV. This is referred to as Model 2, shown in equation 4.2.

$$\begin{aligned} \text{Final Infection}_i = &\alpha + f_1(CC\_i) + f_2(Model_i) + f_3(NetParameter_i) + \\ &f_4(ModelParameter_i) + f_5(CC\_i, NetParameter_i) + \\ &f_6(CC\_i, ModelParameter_i) + \varepsilon_i \end{aligned} \tag{4.2}$$

Furthermore, we investigated the influence of Network Type, as opposed to Model Type, on Non-Zero NPV (referred to as Model 3) as shown in equation 4.3.

$$\begin{aligned} \text{NPV}_i = &\alpha + f_1(CC\_i) + f_2(Network_i) + f_3(NetParameter_i) + \\ &f_4(ModelParameter_i) + f_5(CC\_i, NetParameter_i) + \\ &f_6(CC\_i, ModelParameter_i) + \varepsilon_i \end{aligned} \tag{4.3}$$

We also examined how Network Type influences the Final Infection within the same context, represented by Model 4 as shown in equation 4.4.

$$\begin{aligned}\text{Final Infection}_i = &\alpha + f_1(CC\_i) + f_2(Network_i) + f_3(NetParameter_i) + \\ &f_4(ModelParameter_i) + f_5(CC\_i, NetParameter_i) + \\ &f_6(CC\_i, ModelParameter_i) + \varepsilon_i\end{aligned} \tag{4.4}$$

The model selection, regarding the goodness of fit, is conducted through the Analysis of Deviance procedure using the Likelihood Ratio Test (LRT). We choose a larger model with more predictors only if the p-value for the Analysis of Deviance is less than the significance level of 0.05.

## Results

We fit the model described in Equation 4.1, and the results for Model 1 are shown in Table 4.1. Under our setting, on average, changes in both CC and Network Parameter significantly affect the Non-Zero NPV when other variables in Equation 4.1 are fixed. As shown in the top-left corner of Figure 4.6, a major negative contribution to the Non-Zero NPV from CC occurs at small values between $[0, 0.17]$ and at higher values between $[0.35, 0.45]$. The maximum negative change toward NPV is as large as $-450$ at the minimum value of CC (0.029) and $-500$ at the maximum CC value (0.452). The positive effect of CC arises between $[0.17, 0.35]$. There is a slight decrease between $[0.2, 0.32]$, and the association turns positive again between $[0.25, 0.32]$.

Moreover, the changes of the smooth interaction terms between CC and Network Parameter, indicated by the top-right plot in Figure 4.6 , as well as CC and Model Parameter, the middle-bottom plot in Figure 4.6, both significantly contribute to the changes in Non-Zero NPV when other vari-

Table 4.1: NPV and mean CC Associations by Model Type in Model 1

| Parametric coefficients | Estimate | Std. Error | t value | Pr (t) |
|---|---|---|---|---|
| (Intercept) | 470.22 | 0.80 | 585.43 | **<0.01** |
| ModelICM | 172.43 | 1.02 | 169.09 | **<0.01** |
| ModelLATM | 63.19 | 1.38 | 45.72 | **<0.01** |
| Approximate significance of smooth terms | edf | Ref.df | F | p-value |
| s(CC) | 4.34 | 5.76 | 4.065 | **<0.01** |
| s(NetParam) | 8.90 | 8.98 | 17.93 | **<0.01** |
| s(CC,NetParam) | 26.90 | 27.00 | 2996.83 | **<0.01** |
| s(ModelParam) | 8.99 | 9.00 | 302.42 | **<0.01** |
| s(CC,ModelParam) | 27.00 | 27.00 | 258.27 | **<0.01** |

ables are fixed. These two terms consider the effect of CC towards Non-Zero NPV, conditional on the value of Network Parameter and Model Parameter.

The maximum positive effect from CC and Network Parameter occurs when CC is between $[0.25, 0.3]$ and Network Parameter is between $[0.22, 0.28]$. The maximum negative effect falls at the minimum values of CC and Network Parameter, as suggested at the bottom-left corner in plot 3 from Figure 4.6, indicated by the dark blue color.

The maximum positive effect of the CC and Model parameter interaction term, indicated by the bright yellow color, occurs when CC is at both its minimum and maximum values while the Model Parameter ranges from $[0.5, 0.6]$. The maximum negative effect largely depends on the Model Parameter, either at its maximum value or minimum value, as shown in plot 5 from Figure 4.6

In addition, the changes in NPV differ based on the model type, as shown in the bottom-right corner in Figure 4.6. Specifically, on average, model ICM can increase the Non-Zero NPVs by almost 172 units compared to the reference model LFTM when other variables are fixed. The amount of changes for model LATM is even greater, around 73 units.

The same conclusion as Figure 4.4 is also observed for the response Final Infection, as shown in Figure 4.8. When the filter condition for NPV is applied to be greater than 0, there are almost no

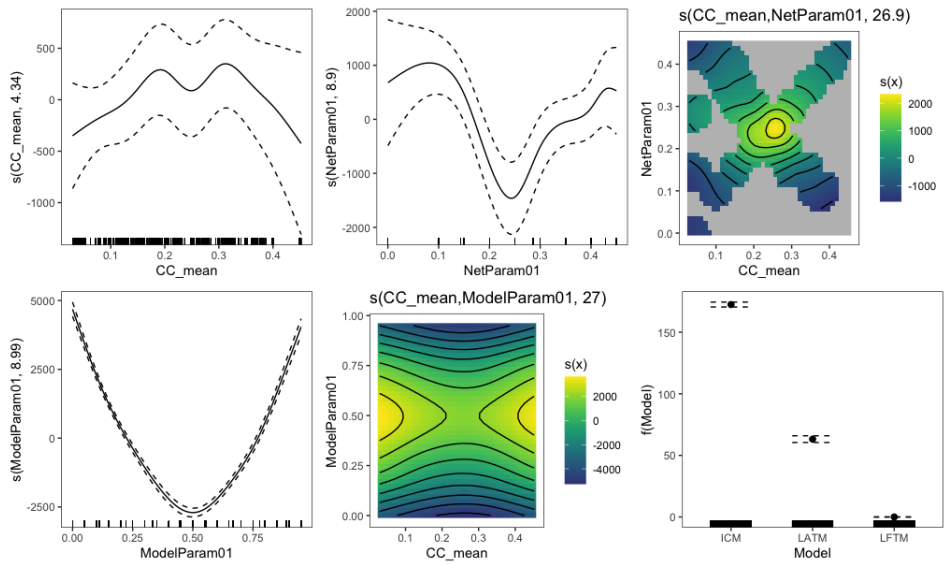



Figure 4.6: GAM Plots for Model 1

Table 4.2: Final Infection and mean CC Associations by Model Type in Model 2

| Parametric coefficients | Estimate | Std. Error | t value | Pr (t) |
|---|---|---|---|---|
| (Intercept) | 0.51 | <0.01 | 562.95 | **<0.01** |
| ModelICM | 0.17 | <0.01 | 144.36 | **<0.01** |
| ModelLATM | 0.07 | <0.01 | 43.76 | **<0.01** |
| Approximate significance of smooth terms | edf | Ref.df | F | p-value |
| s(CC) | 4.69 | 5.82 | 2.01 | 0.06 |
| s(NetParam) | 6.05 | 7.28 | 4.29 | **<0.01** |
| s(CC,NetParam) | 26.21 | 27.00 | 1626.62 | **<0.01** |
| s(ModelParam) | 8.99 | 9.00 | 245.93 | **<0.01** |
| s(CC,ModelParam) | 27.00 | 27.00 | 272.80 | **<0.01** |

average differences between model types and network types.

We also conducted the same analysis with Final Infection as the response, using the smooth term CC as one of the variables of interest (Model 2). Under the same settings, every term, other than the smooth Network Parameter term, is significant towards the Final Infection, as shown in Table 4.2. Model types are significant at the $\alpha = 0.05$ level; however, the magnitude of the changes towards Final Infection between models is much smaller compared to the responses of Non-Zero NPV due to scale differences.

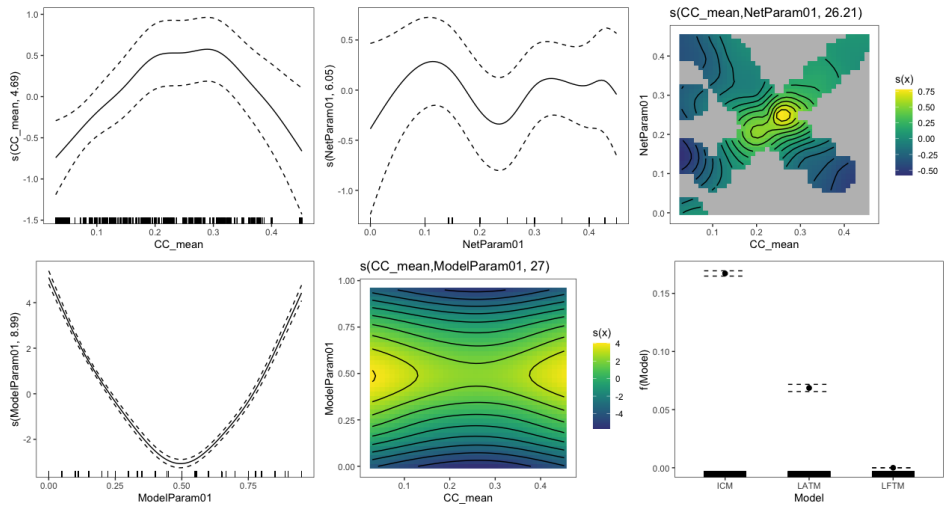



Figure 4.7: GAM Plots for Model 2

We can observe that the smooth effect of CC in Model 2 has a similar shape as in Model 1, as shown in the top left plot of Figure 4.7. The maximum negative effect falls at the minimum values of CC and the maximum values of the Network Parameter, as suggested at the top left and bottom right corners in plot 3 from Figure 4.7. The maximum positive effect of the CC, Model parameter interaction term is when CC and the Model Parameter are at medium value ranges, specifically between 0.25 and 0.3, as indicated by the bright yellow color in the middle. The same trend for the interaction term between CC and Model parameter in Model 1 is observed in Model 2. The maximum positive effect is at either the minimum or the maximum value of CC, as indicated by the bright yellow color in the middle bottom plot in Figure 4.7. The maximum negative effect is at either the minimum or the maximum value of Model Parameter, as indicated by the dark blue color. When compared with the reference Model LFTM, Model ICM can increase the Non-Zero NPV by 0.17 units when all other variables are fixed, and that number increases to 0.29 for Model LATM, as also indicated in the bottom right corner plot in Figure 4.7.

To check the model assumptions, we inspected the residuals for both Table 4.1 and Table 4.2, which appeared to follow a relatively normal distribution based on the histograms and QQ-plots,

as seen in Figure 4.9 and 4.10. At the same time, the meaningful associations discovered in this study can provide insights into NPV and the processes of Final Infection.

In addition to the two analyses mentioned above, we fitted the same two models using Network Type instead of Model Type. Mean CC, our point of interest, is significant at the 0.05 level in both Model 3 and Model 4, shows a similar smoothing effect on both responses but with less curvature in the middle range. More details for Model 3 can be found in Table 4.5, Figure 4.11, and Model 4 results are available in Table 4.6 and Figure 4.12.

Table 4.3 and Table 4.4 below present two simpler models (as shown in equation 4.5 and equation 4.6) in which the predictors include only Model Types, CC, and the interactions between Model Types and CC.

$$\text{NPV}_i = \alpha + f_1(CC\_i) + f_2(Model_i) + f_3(CC\_i, Model_i) + \varepsilon_i \tag{4.5}$$

$$\text{Final Infection}_i = \alpha + f_1(CC\_i) + f_2(Model_i) + f_3(CC\_i, Model_i) + \varepsilon_i \tag{4.6}$$

This is designed to confirm the first hypothesis that, without all other predictors, the effect of CC on both NPV and Final Infection depends on the Model Types. Specifically, without Model Parameter and Network Parameters, a larger CC leads to a bigger increase in both NPV and Final Infection, as shown in plot 1 (top left) in Figure 4.13. Conditioned on the model types, CC affects NPV negatively, as shown in plots 2, 3, and 4 in Figure 4.13. However, the negative impact is consistent across the entire CC interval for Model ICM. Nevertheless, there is a positive trend for Model LFTM when CC is greater than 0.4, and the same applies for Model LATM when CC is

less than 0.1. Similar observations can be made for Final Infection, as shown in Figure 4.14.

Table 4.3: NPV and mean CC Associations by Model Type in Model 1(a)

| Parametric coefficients | Estimate | Std. Error | t value | Pr (t) |
|---|---|---|---|---|
| (Intercept) | 416.45 | 0.92 | 452.40 | <0.01 |
| ModelICM | 233.67 | 1.16 | 201.90 | <0.01 |
| ModelLATM | 202.26 | 1.50 | 135.10 | <0.01 |
| Approximate significance of smooth terms | edf | Ref.df | F | p-value |
| s(CC) | 8.63 | 8.74 | 489.61 | <0.01 |
| s(CC):ModelLFTM | 8.04 | 8.60 | 65.38 | <0.01 |
| s(CC):ModelICM | 0.75 | 0.75 | 178.74 | <0.01 |
| s(CC):ModelLATM | 8.68 | 8.75 | 258.90 | <0.01 |

Table 4.4: Final Infection and mean CC Associations by Model Type in Model 2(a)

| Parametric coefficients | Estimate | Std. Error | t value | Pr (t) |
|---|---|---|---|---|
| (Intercept) | 0.47 | $< 0.01$ | 476.90 | <0.01 |
| ModelICM | 0.21 | $< 0.01$ | 172.50 | <0.01 |
| ModelLATM | 0.18 | $< 0.01$ | 114.40 | <0.01 |
| Approximate significance of smooth terms | edf | Ref.df | F | p-value |
| s(CC) | 8.60 | 8.73 | 351.81 | <0.01 |
| s(CC):ModelLFTM | 7.88 | 8.53 | 61.79 | <0.01 |
| s(CC):ModelICM | 0.75 | 0.75 | 242.65 | <0.01 |
| s(CC):ModelLATM | 8.72 | 8.75 | 157.40 | <0.01 |

Table 4.5: NPV and mean CC Associations by Network Type in Model 3

| Parametric coefficients | Estimate | Std. Error | t value | pr (t) |
|---|---|---|---|---|
| (Intercept) | 399.81 | 7.56 | 52.92 | **<0.01** |
| NetTypeBARABASI | 164.78 | 23.36 | 7.054 | **<0.01** |
| NetTypeRANDOM | 381.07 | 15.60 | 24.44 | **<0.01** |
| Approximate significance of smooth terms | edf | Ref.df | F | p-value |
| s(CC_mean) | 6.29 | 7.37 | 10.29 | **<0.01** |
| s(NetParam) | 1.00 | 1.00 | 0.49 | 0.49 |
| s(CC_mean,NetParam) | 25.41 | 27.00 | 9.15 | **<0.01** |
| s(ModelParam) | 9.00 | 9.000 | 227.49 | **<0.01** |
| s(CC_mean,ModelParam) | 27.00 | 27.00 | 247.92 | **<0.01** |

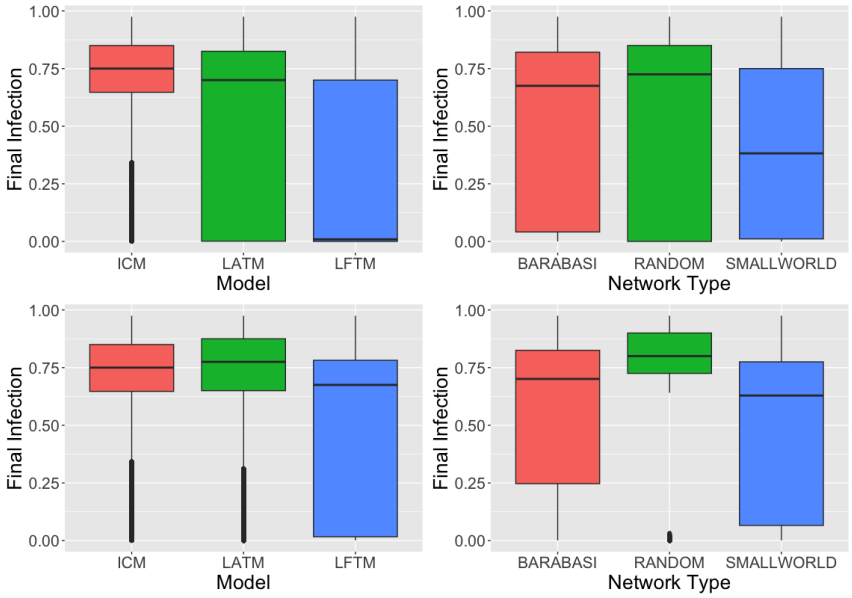



Figure 4.8: Top: All NPV, Bottom: NPV>0

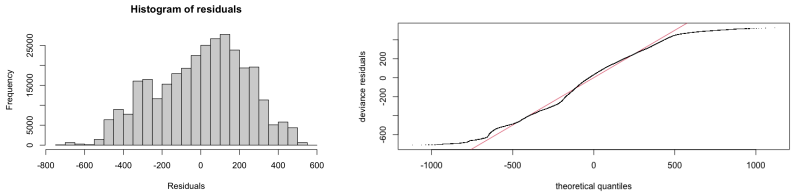



Figure 4.9: Residual and QQ-plots for Model 1

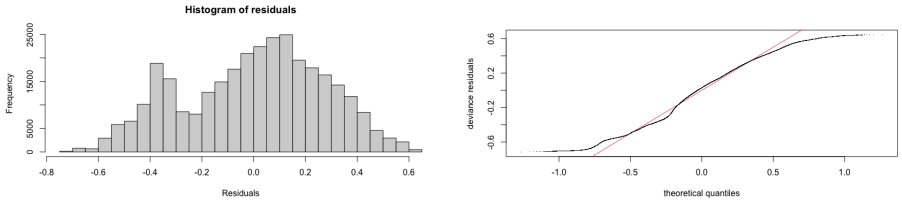



Figure 4.10: Residual and QQ-plots for Model 2

## Discussion

In this study, we examined the impact of CC on both NPV and Final Infection, adjusting for the three Model Types and three Network Types. Under our conditions, CC significantly influenced

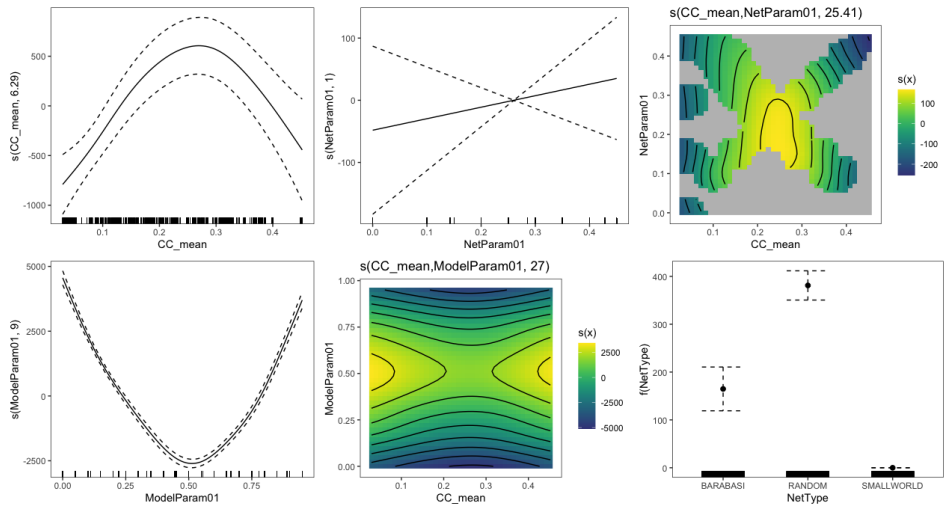



Figure 4.11: GAM Plots for Model 3

Table 4.6: Final Infection and mean CC Associations by Network Type in Model 4

| Parametric coefficients | Estimate | Std. Error | t value | Pr (t) |
|---|---|---|---|---|
| (Intercept) | 0.46 | $< 0.01$ | 176.53 | **<0.01** |
| NetTypeBARABASI | 0.16 | 0.01 | 15.77 | **<0.01** |
| NetTypeRANDOM | 0.33 | $< 0.01$ | 100.38 | **<0.01** |
| Approximate significance of smooth terms | edf | Ref.df | F | p-value |
| s(CC_mean) | 8.93 | 8.97 | 59.95 | **<0.01** |
| s(NetParam) | 1.00 | 1.00 | 20.36 | **<0.01** |
| s(CC_mean,NetParam) | 9.71 | 27.00 | 9.35 | **<0.01** |
| s(ModelParam) | 9.00 | 9.00 | 206.01 | **<0.01** |
| s(CC_mean,ModelParam) | 27.00 | 27.00 | 262.74 | **<0.01** |

Non-Zero NPV, with a more negative effect when CC was either small or large. In the middle range of CC, a positive effect was observed, and the direction of the effect fluctuated around the median value. However, the CC effect on Final Infection, under the same conditions as the previous model, was negatively confounded by Model Types. The same parabolic-shaped smoothing effect observed in Model 1 was also present in Model 2, with the maximum positive effect of CC and Parameters occurring in the middle range for both models.

Our results support the hypothesis that the CC's effect on both NPV and Final Infection is conditionally dependent on the Model Parameter and Network Parameters at the 0.05 significance level.

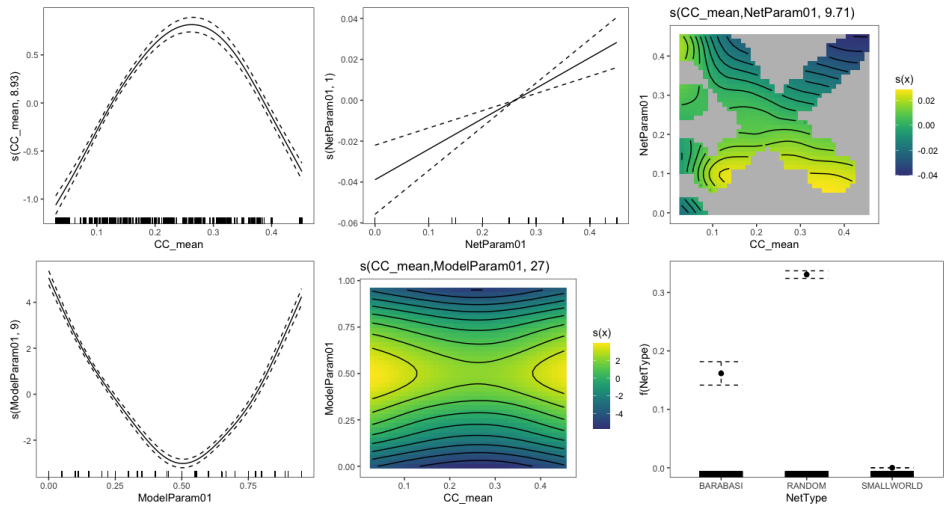



Figure 4.12: GAM Plots for Model 4

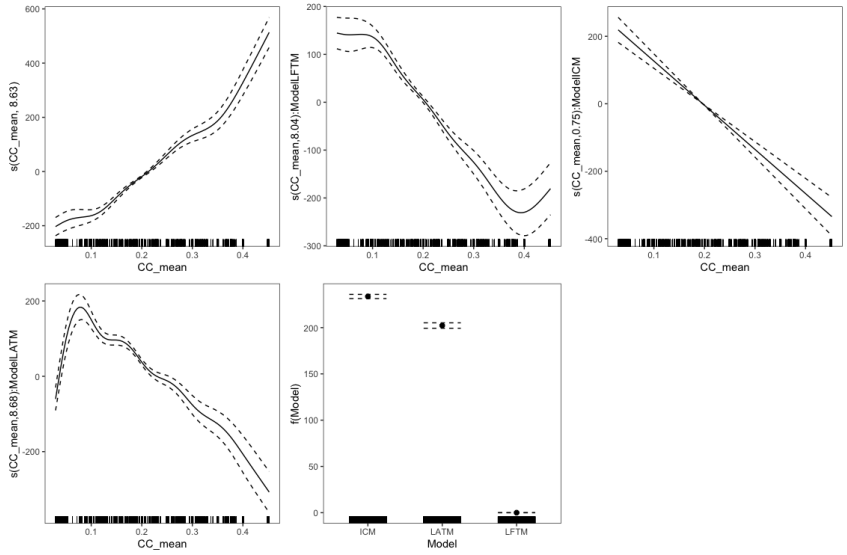



Figure 4.13: GAM Plots for Model 1 (a)

The effect of CC on NPV is dependent on the Model Type, as shown in Table 4.3 and Figure 4.13. Similarly, this dependence on Model Type applies to Final Infection, as indicated in Table 4.4 and Figure 4.14.

Moreover, the insights from this study explain the relationship between NPV (and final infection), analyzed through CC, DOI model, and network type. This comprehensive analysis enhances our

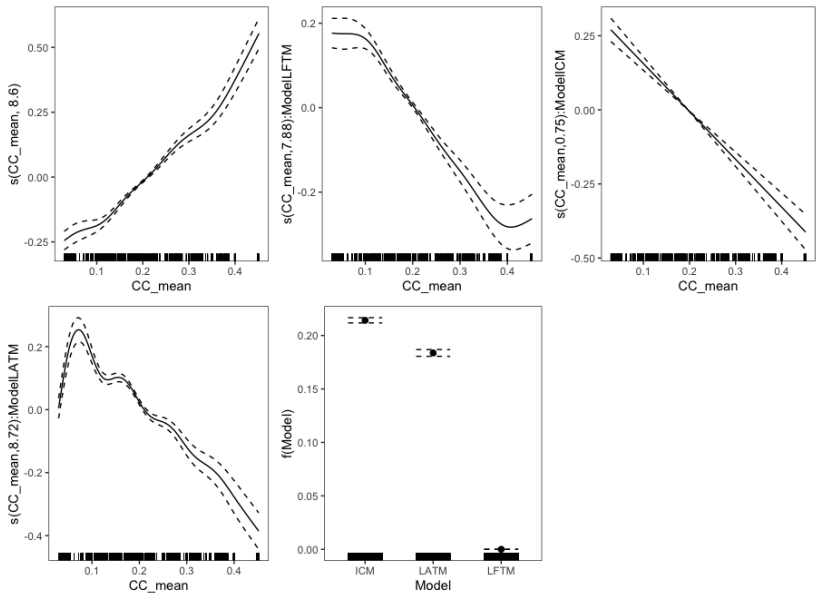



Figure 4.14: GAM Plots for Model 2 (a)

understanding of how clustering (existence of communities) affects the information diffusion dynamics concerning network types and diffusion models.

# CHAPTER 5: MEASURING INFLUENCE IN ONLINE SOCIAL NETWORKS

Online social networks such as Twitter have become a platform enabling users to share information and influence public opinion. Understanding the dynamics of user interaction on Twitter, such as tweeting, retweeting, and mentioning, is essential to understanding how information spreads and what factors contribute to an individual's influence on the platform.

This part of our research aims to propose a meaningful method of identifying influential users in platforms such as Twitter. We propose a methodology that utilizes Transfer-entropy as a way of measuring importance of users in the information diffusion space. Then we develop a methodology for comparing various influence measures of OSNs by using the user adoption as a reference. such as measured by various methods such as the number of received retweets, number of received mentions, and the number of posted Tweets. Further, we focus on the fact that influence changes over time. Therefore, influence measurement of a user should be changing over time. For example, the number of received retweets of a user would be different each day.

## Methodology

We hypothesize that time periods in which certain users become influential, their influence correlate with the growth of the volume of tweets, suggesting that the influential power of those certain users have amplified the information cascade. One way to measure influence is the number of retweets a certain user receives. At each time interval, a user receives a certain total number of retweets, which shows the user's influence. We refer to this measurement as Received Retweet Count (RRC) in this text. RRC was used as a influence measurement in previous work we dis-

cussed chapter 2 [12, 3]. Another way that influence could be measured is by using TE as used by [42]. However, in their work they calculated TE of the whole dataset at once. In our approach we want to have a TE value for each user at each time interval. Therefore, we should calculate TE continuously as we move forward through time, step by step through time intervals.

### *Data Collection*

Initially, we gathered Tweets[1] discussing wildfires directly through the Twitter API, collecting data from the entire Twitter platform from January 1st to April 1st, 2022 (120 days of data). The dataset we collected ended up having approximately 3500 unique users. Afterward, we collected all the Tweets these users created in the same period (January 1st to April 1st, 2022). This data set is a complete dataset that allows us to track all of the Twitter activity of these users during the mentioned period. From this extensive dataset, we selected three highly shared hashtags that cover distinct topics for creating three datasets for our analysis: #ukraine (56430 tweets), #climateaction (23629 tweets), and #covid19 (35109 tweets).

### *Transfer-entropy based measurement*

Transfer-entropy (TE), introduced by Schreiber [41], is a information theoretical measurement based on Shannon entropy [43]. Given two random processes, TE quantifies how much uncertainty in predicting the next state of one process is reduced by incorporating the histories of both processes. In other words, TE shows weather it becomes more easy to predict a target process when we use the history of both the target process and a source process. The value of TE obtained is in the range $[0, 1]$ and it shows the strength of the relationship. TE is not-commutative, in other

[1]The term Tweets here is used loosely to represent all forms of Twitter posts such as tweets, retweets, quoted tweets, and replies.

words TE is directional.

Let $X_t$ and $Y_t$ be the random processes ($t \in \mathbb{N}$). Then Transfer-entropy from $X$ to $Y$, $TE_{X \to Y}$ is calculated as given in equation 5.1 [41].

$$TE_{X \to Y} = \sum P(Y_{t+1}, Y_t^{(k)}, X_t^{(l)}) \log \frac{P(Y_{t+1} | Y_t^{(k)}, X_t^{(l)})}{P(Y_{t+1} | Y_t^{(k)})}, \tag{5.1}$$

$Y_t^{(k)}$ denotes a sequence of history of length $k$ of the $Y$ starting from $t^{th}$ time-step as: $(Y_t, Y_{t-1}, ..., Y_{t-k+1})$. $k$ and $l$ are history lengths of $Y$ and $X$ respectively. In this study we keep the history lengths equal, and therefore $k = l$. For the purpose of our study we rewrite the same equation as follows, denoting the $k$ as a parameter:

$$TE_{X \to Y}(k) = \sum P(Y_{t+1}, Y_t^{(k)}, X_t^{(k)}) \log \frac{P(Y_{t+1} | Y_t^{(k)}, X_t^{(k)})}{P(Y_{t+1} | Y_t^{(k)})}, \tag{5.2}$$

For each user in a dataset, we begin by extracting the activity time series, identifying all timestamps when the user posted a Tweet. Next, we perform a resampling at a frequency $f$, dividing the entire time span of the dataset into bins of size $f$. This results in a resampled time series of user activity, where each time-step is marked with a 1 if the user tweeted during that interval, and a 0 if not. In this study we have chosen the resampling frequency as $f = 1$ Day. This binary time-series of each user is used for calculation of TE by applying it to the equation 5.2.

The dataset we have gathered is 120 days long. Therefore, with 1 Day time intervals, we have 120 steps. We calculate TE between all ordered pairs of users for time-step $t$ by using a growing window which always starts at the time-step 0 and ends at time-step $t$. Therefore, the amount of influence from user A to user B at time $t$ is $TE_{A \to B}(t)$. We calculate the influence measurement of

user A at time-step $t$ by taking the total outgoing TE, which we denote as $TTE_A(t)$. In the equation 5.3 $B$ is all potential neighbors of $A$ which is all users except $A$. This calculates the total of TE values for all outgoing edges of node A.

$$TTE_A(t) = \sum_{allB} TE_{A \to B}(t) \tag{5.3}$$

With the above equation 5.3, we calculate the TE based influence measurement called the Total Outgoing Transfer Entropy (TTE). For the rest of the text, we denote the received retweets count as RRC which we mentioned as the traditional method of measuring influence.

For comparing the TTE and RRC we use volume curve (which is the volume of Tweets over time) and the adoption curve (which is a curver overtime representing the number of new users joined to the conversation at the given time interval).

*Statistical Analysis*

Adoption curve reflect the impact of social influencers and predicting the adoption curve can guide the informative decisions [8]. From our Modeling Information Pathways (MIPs) perspective, adoption curve indicates the information diffusion when it refers to the topics discussed on the social media platforms. MIPs project aims to understand the information flow that promote early detection of the influential messages, and accurately predict the adoption curve can help to guide the early stage identification.

We are interested in identifying an effective measurement that can accurately predict the adoption curve based on observations from a set of individuals. Specifically, we aim to compare the following two measurements: Total Transfer Entropy (TTE), and Received Retweet Count (RRC).

For example, the scatter plot on the left of Figure 5.1 shows the raw adoption observations about the conversations related to the #ClimateAction, with the red line representing the smoothed adoption curve. On the right of Figure 5.1, the total transfer entropy over time is displayed for 487 users over a 120-day period since the eruption of the event. We are investigating the potential of using TTE curves to predict the adoption curve.

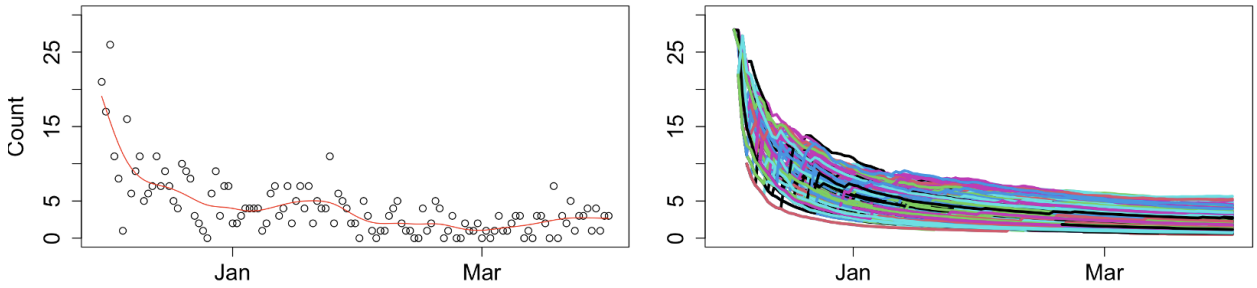



Figure 5.1: Adoption Curve and Total Transfer Entrophy Over Time of #ClimateAction

As shown on the left of Figure 5.2, the mean TTE curve closely resembles the trend of the adoption curve over time. Predicting using users who reveal the least amount of departure from the mean curve is promising, as indicated by the right graph in Figure 5.2.

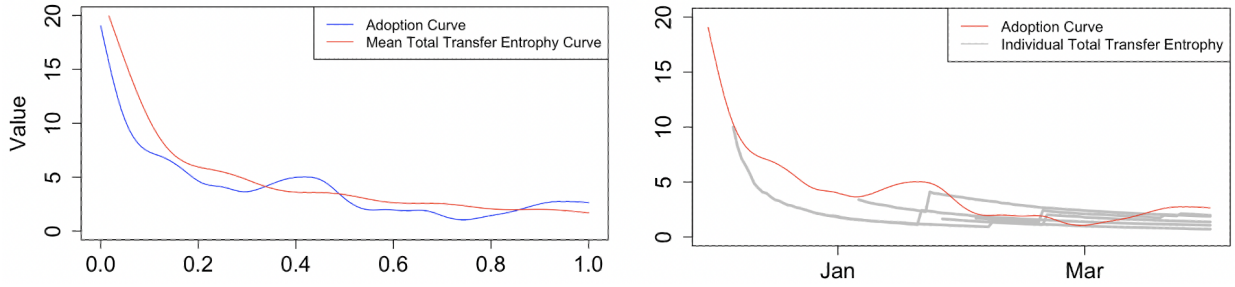



Figure 5.2: Adoption Curve Prediction Using Mean TTE Curve of #ClimateAction

However, the adoption curve displays distinct characteristics depending on the information cycle as well as the influencers. For example, in Figure 5.3 below, the same 487 users are discussing Ukraine, a different topic from the Climateaction. Unlike the topic of the Climateaction, both the

adoption curve and volume counts for the topic Ukraine peaked near the end of the study time. The topic-dependent adoption curve proves to be challenging to accurately model, and it is critical to identify proper metrics that can capture the variability of the adoption curve regardless of the topics.

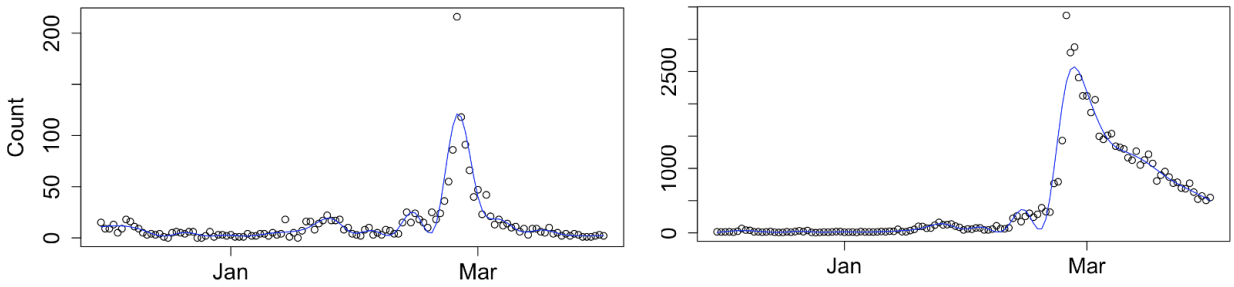



Figure 5.3: Adoption Curve and Volume Curve of #Ukraine

We also explore the possible association between RRC and the adoption curve. In general, the direction of movement from RRC does not align well with the shift of the adoption curve compared to TTE. However, a potential non-direct association might be possible to connect the RRC with the adoption curve, and that is one of the future research directions. In this paper, we juxtapose the effective predictions from TTE and RRC in terms of L2 distance [36], which is used to evaluate the closeness of the two curves. Specifically, denote the $(Y_1, t_{1j}), \cdots, (Y_i, t_{ij})$ is the observed adoption values over a compact interval $\mathscr{T}$ where $i = 1, \cdots, n, j = 1, \cdots, m$, we smooth the discretely observed adoption observations and apply a roughness penalty. Then, we calculate the mean TTE curve from individual TTE over time, and the mean RRC curve from individual RRC curves. The $L^2$-distance between the adoption curve $Z(t)$ and the mean TTE curve $\widetilde{Z}(t)$ or mean RRC curve $Z^{*]}(t)$ is defined as:

$$d = \int_0^1 \{\widehat{Z}(t) - \widetilde{Z}(t)\}^2 dt.$$

We aim to compare the closeness of the adoption curve prediction using the mean curve of the TTE as well as the mean curve of RRC. Additionally, we are interested in identifying a group of

important users from the individuals in the sample that can explain the variability of the adoption curve.

In order to smooth the discrete adoption curve with a dense sampling design, parameter estimation is generally estimated through penalized cubic spline functions [55]. In particular, we consider the smoothed adoption curves can be estimated through the linear combination of basis functions where the only unknown is the coefficients of the basis functions. Let $B_m$ is the $M$-dimensional vector with the $m$th element equal to $B_m(t_{ij})$ and $\beta$ is the $M$-dimensional vector with the $m$th element equal to $\beta_m$, we consider the adoption curve such that $Z(t) = \sum_{m=1}^{M} \beta_m B_m(t)$. To accommodate a large class of adoption curves, the dimension of the basis $M$ is chosen large and the parameters $\beta$ are estimated using a penalized criterion to penalize overfitting [36, 13]. Define the penalized log-likelihood function by

$$p\ell(\beta) = -2\ell_i(\beta_i) + \lambda \beta S \beta \tag{5.4}$$

where $S$ is an $M \times M$ known penalty matrix, representing the smoothness for the adoption curve, and $\lambda$ is a penalty parameter that ensures the smoothness of the adoption curve is captured and to avoid the overfitting with too much curvature [54]. We use the common second order penalty with the $(m,m')$th element given by $\int \{B_m''(t) B_{m'}''(t)\} dt$, where $B_m''(t)$ is the second derivative of $B_m(t)$, for $m = 1, \ldots, M$. The estimates $\widehat{\beta}$ are obtained by minimizing the penalized criterion $p\ell_i(\beta)$, and the optimal value of the penalty parameter $\lambda$ is selected through Generalized Cross Validation (GCV) using Maximum Likelihood-based approach [56].

For the mean TTE curve and the mean RRC curve as well as the covariance estimation to identify the individual variabilities, we use the fast Covariance Estimation for High-dimensional Functional Data for dense grid introduced in [57] to perform the functional principal component analysis (FPCA), where a fast implementation of the sandwich smoother is adopted and a two-step

procedure that first applies singular value decomposition to the data matrix and then smoothes the eigenvectors. FPCA allows us to estimate the mean function that quantifies the overall behavior of either TTE or RRC as well as the covariance function of each which evaluate the variability of each individual from the mean curve.

Specifically, let the FPCA be

$$\widetilde{Z}(t) = \widehat{\mu}(t) + \sum_{k=1}^{K} \widehat{\xi}\widehat{\phi}_k(t)$$

using a large percentage of explained variance (PVE) such as 95%, where $\widehat{\mu}(\cdot)$ is the smooth estimated mean function that can be used to calculate the L-2 distance between the adoption curve, $\widehat{\phi}_k(\cdot)$ is the $k$th estimated eigenfunction and $\widehat{\xi}_k$ is the predicted FPC score. In our case, we selected the first FPCA scores that accounts for at least 87% of the varibility to identify the group of representative users. To be specific, the high varibility users is defined as the top users who have the largest $\widehat{\xi}_1$ and the low varibility users is defined as the last few users who have the smallest $\widehat{\xi}_1$.

## Results

We investigate three topics, Wildfire, Ukraine War as well as Covid19, for the same group of individuals. These individuals are selected as follows:

For each topic, we estimated the adoption curve, using the mean TTE curve and mean RRC curve to predict the adoption curve, calculate the L2 distance bwtween the mean curve and the adoption curve, then we selected a subset of individuals who might have the potential to predict the adoption curve without compromising too much information. The selection criteria includes the following:

1. The high varability users

2. The low variability users

3. Top 10% TTE sum for the first 14 days of period since the event

4. First 10% of the individuals

The L2 distance is calculated using the overall mean curve and the mean curve from the above four criterias. We compare the results of each setting and below is the summary figure of each topic.

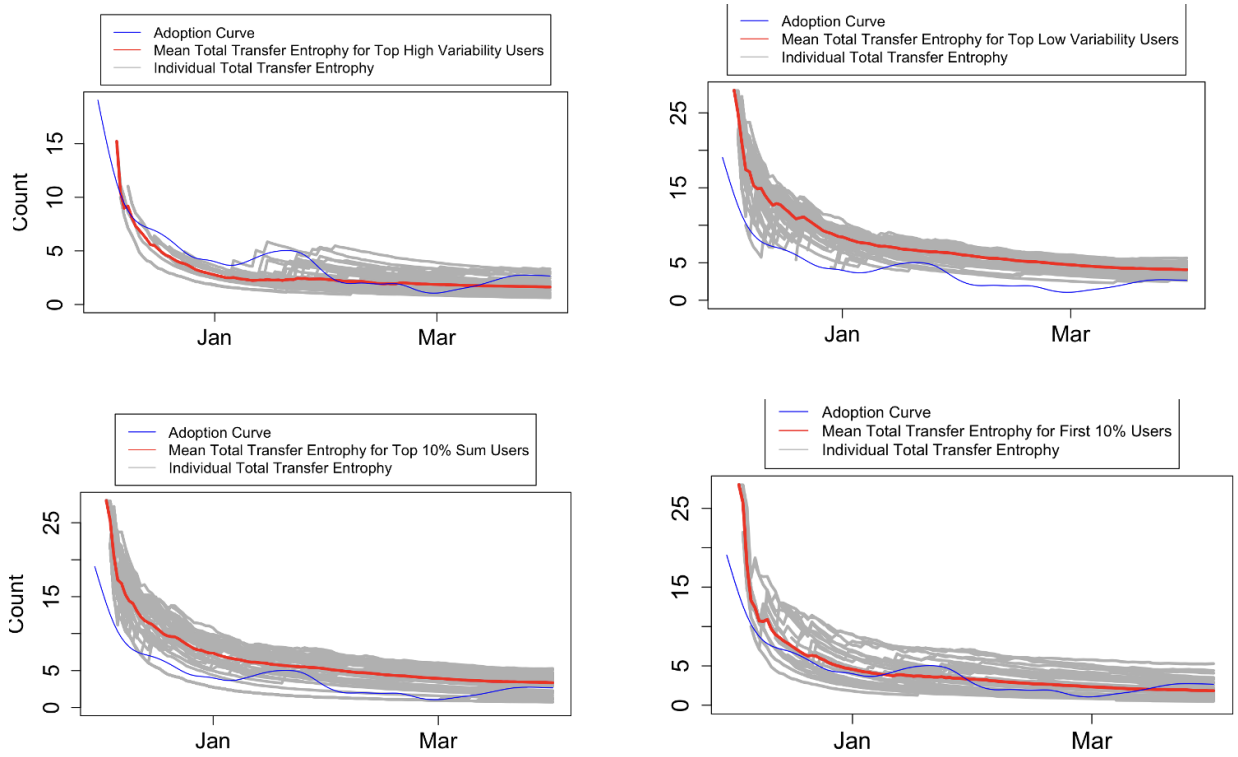



Figure 5.4: Climateaction TTE prediction use selected individuals

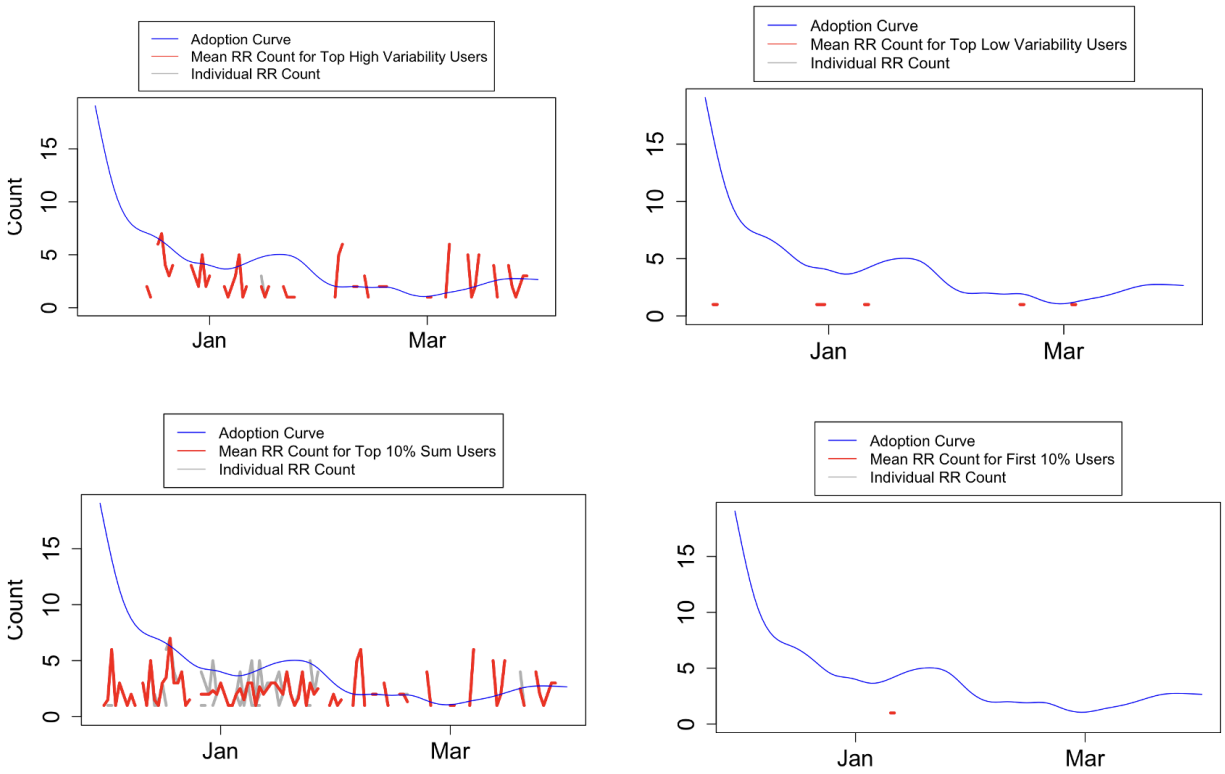



Figure 5.5: Climateaction RRC prediction use selected individuals

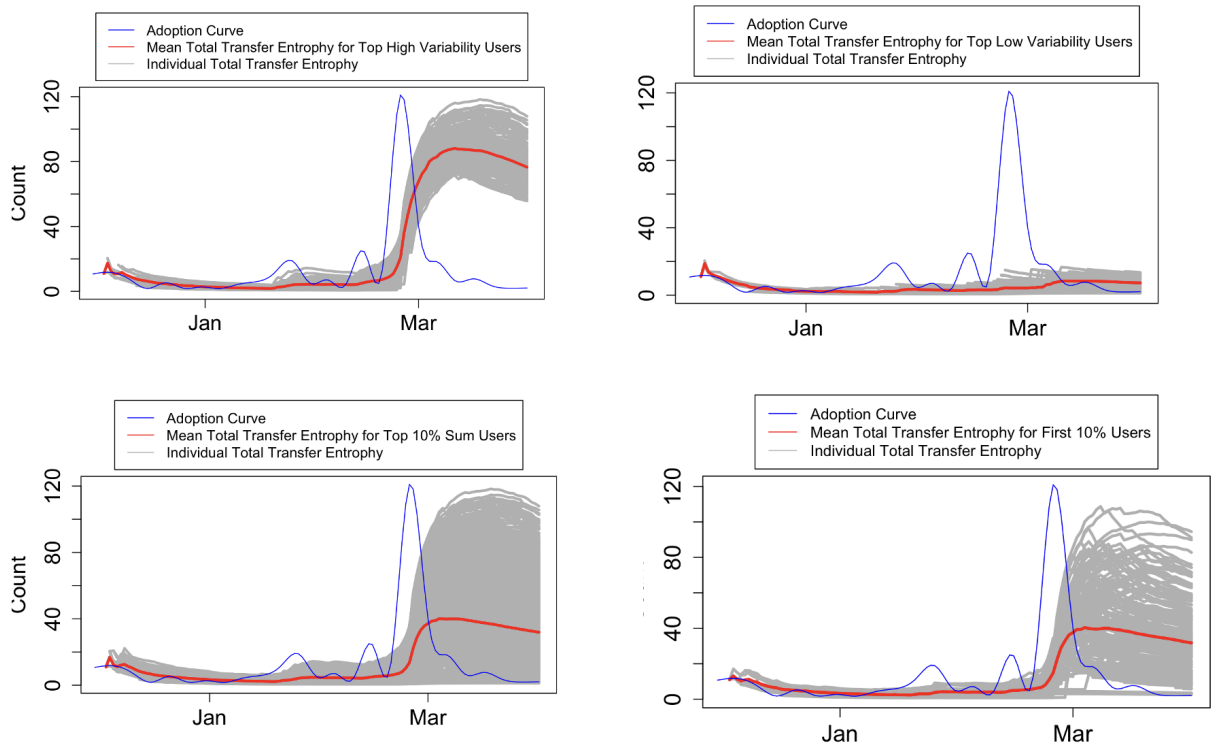



Figure 5.6: Ukraine TTE prediction use selected individuals

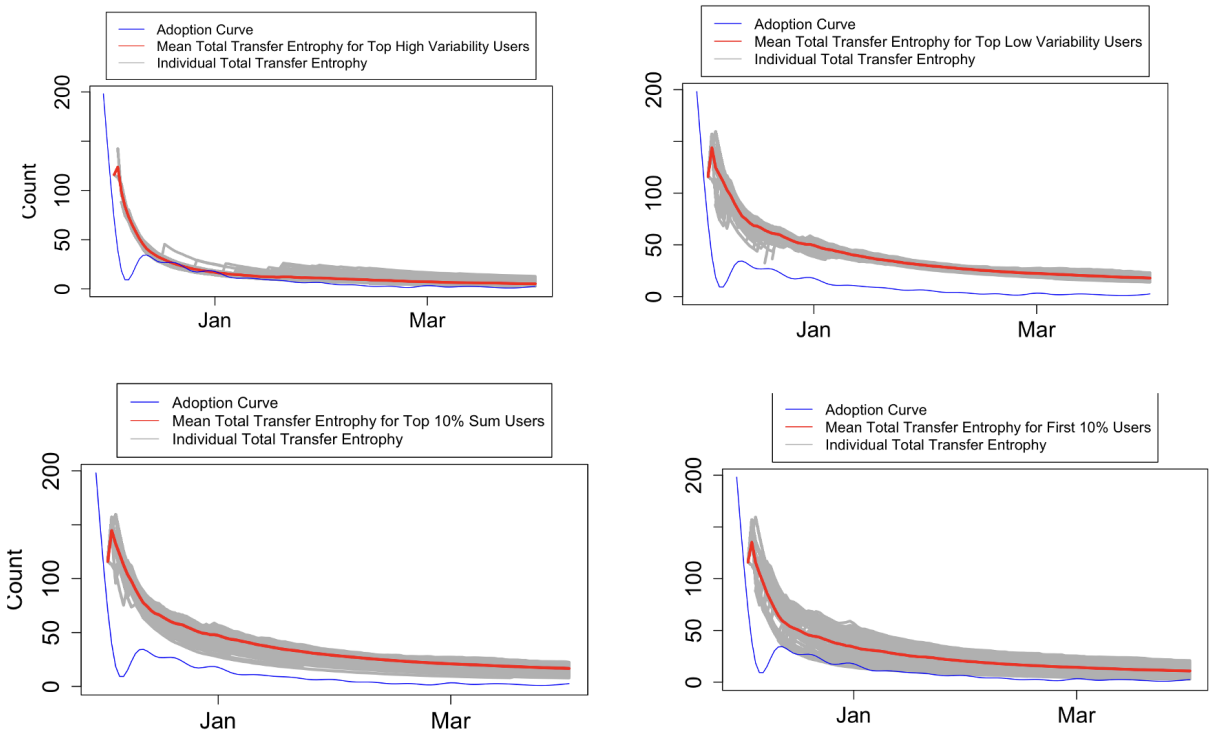



Figure 5.7: Covid19 TTE prediction use selected individuals

Overall, TTE predicts the adoption curve relatively better compared to the RRC prediction under our scenario due to the fact that TTE curve in general follows the trend of the adoption curve in both direction and the raw value. Explore the association between the RRC as well as the adoption curve can be an interesting next step.

Table 5.1 is the L2 distance between the adoption curve as well as the mean curves from each setting. The smaller the value is, the better the prediction is. We can see that the L2 distance between the adoption curvea and mean TTE curves are smaller compared to the RRC across all topics and all scenarios. Topic Climateaction and Covid19 peak in the begining and they also both have smaller L2 TTE distances compared to Ukraine. In some cases the RRC value looks better than the TTE value, the actual curve given by RRC is not useful (curve is closer to 0 in all times, thus the L2 distance better than the TTE) as shown in the respective figures. Therefore, we can

conclude from this analysis that TTE is a better measurement than RRC when comparing their strengths in directly estimating the user adoption curve.

Table 5.1: L2 distance between the adoption curve and mean curves

| | Climateaction | | Ukraine | | Covid19 | |
|---|---|---|---|---|---|---|
| L2 between adoption curve | TTE | RRC | TTE | RRC | TTE | RRC |
| Overall mean curve | 1.71 | 3.33 | 24.46 | 25.13 | 22.40 | 17.54 |
| High varability users | 2.44 | 3.89 | 41.08 | 23.68 | 15.11 | 18.78 |
| Low variability users | 3.98 | 4.55 | 24.91 | 26.49 | 33.67 | 17.12 |
| Top 10 TTE Sum | 3.24 | 3.47 | 24.90 | 25.00 | 33.95 | 17.50 |
| First 10 percent | 1.90 | 4.70 | 24.82 | 26.55 | 23.86 | 18.76 |

## Discussion

We observe that TTE is a smoother curve. The subset curves generated using TTE curves show a closer direct relationship with each dataset's adoption curve. Meanwhile, RRC is more of a discrete curve showing different signals. We were unable to find a direct relationship between RRC and the adoption curve. It could be either that RRC doesn't have a useful relationship with the adoption curve or that the relationship between RRC and the adoption curve is more complex. The direct relationship we have drawn using the RRC and TTE shows that TTE is far more helpful in directly estimating the adoption curve than RRC. This leads us to a new understanding of the validity of the Transfer Entropy-based measure, TTE, in identifying influential actors within online social networks regarding information diffusion and adoption.

Regardless of the results we also see a lot of parameters that could have been adjusted differently such as the sampling frequency $f$, and history length $k$. Potential immediate future work resides in

the area where different parameter settings are explored for RRC.

# CHAPTER 6: CONCLUSIONS

With the aim of comparing conceptually different models under a common framework, we formulated a parametric general model of information diffusion. Using this formulation we identified two important dimensions which led to creating a conceptual framework based on two properties of models of information diffusion: *neighbor knowledge* and *stochasticity*. This framework allowed us to classify existing classical DOI models into mechanistically distinct classes. We compared the dynamics of these conceptually different DOI models on directed scale-free networks in order to identify whether the underlying diffusion characteristics of each model differed.

We found that BRRM and DWM, regardless of having different conceptual designs, converge to a the same final state wherein all reachable nodes are infected. Furthermore, the sensitivity of their parameters for final state is only determined by the network parameters. Meanwhile, their model-specific parameters are able to alter the speed of information diffusion.

Through analysis of linear threshold models (LATM and LFTM) and their stochastic counterparts (SLATM an SLFTM) we learnt how adding stochasticity to a model may not always change the final state. In general, a stochastic version of a model created by appending a probability check to the final step of the existing rule will yield a model that has its final state bounded by the final state of the original model. If the added stochastic step is allowed to run at every time-step, this new stochastic model will be slower than the original one, will take more time-steps to converge, and the sensitivity of the stochastic probability parameter with respect to both $\phi_F$ and NPV will be close to zero. However, the sensitivity to other parameters could change.

We conclude that despite being conceptually similar, DOI models may exhibit significantly different behavior in terms of final state, speed of infection diffusion, and sensitivity to final fraction of spread. Models belonging to the same conceptual class may produce completely different simula-

tion outcomes (e.g., BRRM vs SLFTM, ICM vs DWM, and ICM vs SLATM), while models belonging to different classes may yield similar outcomes (e.g., BRRM vs DWM, LFTM vs SLFTM, and LATM vs SLATM ). This suggests that the behavior of models cannot be solely predicted by their conceptual class, but rather requires a more detailed analysis of their specific dynamics. Hence, it is important to investigate the unique behavioral characteristics of a model (despite the conceptual design) when choosing an appropriate DOI model for a specific application.

The investigation into the impact of clustering coefficient (CC) on NPV and Final Infection unveiled a complex interplay between CC, model types, and network parameters. The observed conditional dependence of CC's effects emphasizes the complex relationship between network topology and diffusion dynamics, highlighting the role that clustering plays in influencing information spread.

In the follow up study we investigate the effect of clustering on the outcomes of three different DOI models (ICM, LATM, and LFTM) under three network structure types (R, SW, and SF). Again, the final fraction of infection ($\phi_F$) and net present value (NPV) were considered as the outcomes of interest. Clustering of each network were measured using the mean clustering coefficient. Through statistical analysis, we confirm that the outcomes of DOI depend on clustering and that the effect of clustering (on both NPV and $\phi_F$) depends on network parameters and model parameters. Moreover, we find statistical evidence that the model type determines the effect of clustering. Our findings are a hint that both DOI model type and clustering are important in determining the final outcome of a DOI run. Furthermore, it's essential to acknowledge that network parameters significantly influence the impact of clustering on diffusion outcomes. We plan to continue this investigation on empirical networks as part of our future work, with the intention of expanding our research in this area.

Lastly, we introduced Total Outgoing Transfer Entropy (TTE) as a Transfer Entropy-based mea-

sure for measuring influence in a Twitter dataset. Comparing TTE versus RRC revealed TTE's ability to estimate adoption curve. This contrast between TTE and RRC offers a fresh perspective on assessing influence and adoption, marking TTE as a more direct and reliable metric for understanding information diffusion trends.

## Future Work

The experimental analysis in this work covers the entire parameter space of all models in an unbiased manner since goal is to compare the outcomes of the models across the parameter space and gain theoretical insights. However, in real-world scenarios, the parameter space of the actual data may follow some probability distribution. Therefore, when applying these models to specific contexts, it is crucial to understand the underlying parameter distribution of that context before conducting a similar experimental analysis. This ensures that the results are more representative of the actual scenario and improves the applicability of the models in real-world situations.

An additional avenue for future research would involve conducting a dis-aggregated analysis of the parameter space to examine its impact on the final outcome of each model. By conducting such an analysis, we can gain deeper insights into the specific contributions and interactions of different parameters in shaping the final outcomes of the models.

In above discussions we have shown how models can have different rates of infection at different stages of the simulation regardless of the number of final infected nodes (e.g., compare how BRRM infects a greater number of nodes within early stages but lags behind compared to another stochastic model such as SLATM after $\phi_F \geq 65\%$). This results indifferent shapes of adoption curves. Studying the patterns of the adoption curves in more detail may be a possible future work in order to classify capabilities of different models. Another possible aspect to be investigated is the depth

and breadth of the subnetwork of infected nodes as a characteristic of models. For instance, do certain properties of the network structure, e.g. degree distribution, make it more likely for a node to become infected in one of these models versus the other ones. An expansion on this would be the notion of "structural virality"[20, 32]. Investigating the relationship between these structures and model infection rates should be investigated in future work.

Finally, there is a lot of possible immediate future work on top of the TTE influence measure. It could be compared against other existing measures such as centrality based measures. Also it is critical to perform the comparison with varying sample rates, history lengths, and also with moving windows.

# APPENDIX A: IRB OUTCOME LETTER

UNIVERSITY OF CENTRAL FLORIDA

**Institutional Review Board**
FWA00000351
IRB00001138, IRB00012110
Office of Research
12201 Research Parkway
Orlando, FL 32826-3246

# NOT HUMAN RESEARCH DETERMINATION

January 17, 2024

Dear Chathura Don Dimungu Arachchige:

On 1/17/2024, the IRB reviewed the following protocol:

| | |
|---|---|
| Type of Review: | Initial Study |
| Title of Study: | Information Diffusion Mechanisms and the Effect of Influential Users within Online Social Networks |
| Investigator: | Chathura Don Dimungu Arachchige |
| IRB ID: | STUDY00006317 |
| Funding: | Name: Defense Advanced Research Projects Agency (DARPA) |
| Documents Reviewed: | • HRP-250 - FORM - Request for NHSR Chathura.docx, Category: IRB Protocol; |

The IRB determined that the proposed activity is not research involving human subjects as defined by DHHS and FDA regulations.

IRB review and approval by this organization is not required. This determination applies only to the activities described in the IRB submission and does not apply should changes outside of administrative ones (study personnel, timelines, etc.) be made. If non-administrative changes are made (design, information collected, instrumentation, funding, etc.) and there are questions about whether these activities are research involving human in which the organization is engaged, please submit a new request to the IRB for a determination by **clicking Create Modification / CR** within the study.

If you have any questions, please contact the UCF IRB at 407-823-2901 or irb@ucf.edu. Please include your project title and IRB number in all correspondence with this office.

Sincerely,

Kristin Badillo
UCF IRB

# APPENDIX B: ADVANCES IN COMPLEX SYSTEMS COPYRIGHT PERMISSION TO REUSE THE AUTHOR PUBLISHED PAPER

# Re: Request for Permission to Reuse Article in PhD Dissertation

Journal ACS <acs@wspc.com>

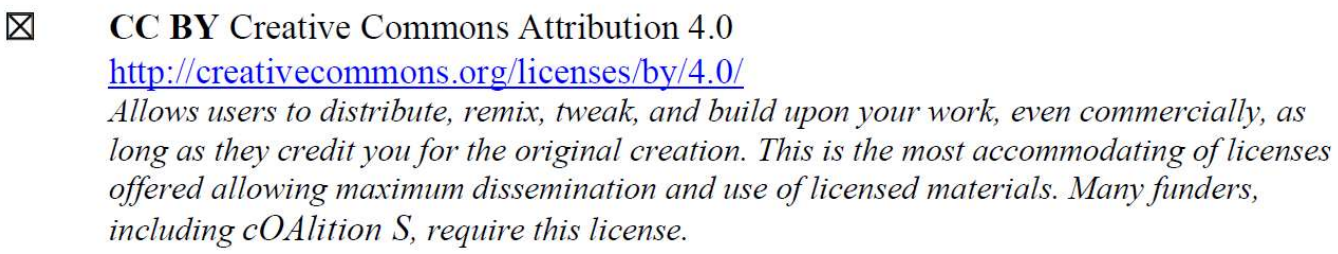
Thu 3/21/2024 11:32 PM

To:Chathura Don Dimungu Arachchige <chathura@ucf.edu>
Cc:karsaim@ceu.edu <karsaim@ceu.edu>

Dear Chathura Jayalath,

I'm sorry I couldn't respond earlier. The OA license that you had opted (imaged below) allows for your re-use in the most maximal and relaxed terms.
You will only need to credit the original version to re-use.
Hope this helps.

Sincerely,
Yong Qi
Editor, World Scientific
on behalf of Advances in Complex Systems






---

**From:** Chathura Don Dimungu Arachchige <chathura@ucf.edu>
**Sent:** Sunday, March 17, 2024 11:18 PM
**To:** Journal ACS <acs@wspc.com>; karsaim@ceu.edu <karsaim@ceu.edu>
**Subject:** Request for Permission to Reuse Article in PhD Dissertation



Dear Editors,

I am reaching out to inquire about the process for obtaining permission to include the article titled "A Generalization of Threshold-Based and Probability-Based Models of Information Diffusion" (https://doi.org/10.1142/S0219525923500054) in my doctoral dissertation. I hold the position of first author for this paper, which was developed as a component of my PhD research. My dissertation is set to be digitally published by the University of Central Florida.

Could you please advise if such reuse is permissible, and if so, what steps should be taken to formally request authorization?

Thank you,
Chathura Jayalath

# LIST OF REFERENCES


[1] Mathilda Åkerlund. The importance of influential users in (re) producing swedish far-right discourse on twitter. *European Journal of Communication*, 35(6):613–628, 2020.

[2] Aris Anagnostopoulos, George Brova, and Evimaria Terzi. Peer and authority pressure in information-propagation models. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 76–91. Springer, 2011.

[3] Ali Sajedi Badashian and Eleni Stroulia. Measuring user influence in github: the million follower fallacy. In *Proceedings of the 3rd International Workshop on CrowdSourcing in Software Engineering*, pages 15–21, 2016.

[4] Albert-László Barabási and Réka Albert. Emergence of scaling in random networks. *science*, 286(5439):509–512, 1999.

[5] Frank M Bass. A new product growth for model consumer durables. *Management science*, 15(5):215–227, 1969.

[6] Jonathan D Bohlmann, Roger J Calantone, and Meng Zhao. The effects of market network heterogeneity on innovation diffusion: An agent-based modeling approach. *Journal of Product Innovation Management*, 27(5):741–760, 2010.

[7] Béla Bollobás, Christian Borgs, Jennifer Chayes, and Oliver Riordan. Directed scale-free graphs. In *Proceedings of the fourteenth annual ACM-SIAM symposium on Discrete algorithms*, pages 132–139. Society for Industrial and Applied Mathematics, 2003.

[8] William A Brock and Steven N Durlauf. Adoption curves and social interactions. *Journal of the European Economic Association*, 8(1):232–251, 2010.

[9] Damon Centola. The spread of behavior in an online social network experiment. *science*, 329(5996):1194–1197, 2010.

[10] Damon Centola, Víctor M. Eguíluz, and Michael W. Macy. Cascade dynamics of complex propagation. *Physica A: Statistical Mechanics and its Applications*, 374(1):449–456, 2007.

[11] Damon Centola and Michael Macy. Complex contagions and the weakness of long ties. *American journal of Sociology*, 113(3):702–734, 2007.

[12] Meeyoung Cha, Hamed Haddadi, Fabricio Benevenuto, and Krishna Gummadi. Measuring user influence in twitter: The million follower fallacy. In *Proceedings of the international AAAI conference on web and social media*, volume 4, pages 10–17, 2010.

[13] Xiaoxia Champon, Ana-Maria Staicu, Anthony Weishampel, Chthura Jayalath, and William Rand. Clustering of categorical valued functional data with application to social media. 2023.

[14] Sebastiano A Delre, Wander Jager, and Marco A Janssen. Diffusion dynamics in small-world networks with heterogeneous consumers. *Computational and Mathematical Organization Theory*, 13(2):185–202, 2007.

[15] Jacques Demongeot, Quentin Griette, and Pierre Magal. Si epidemic model applied to covid-19 data in mainland china. *Royal Society Open Science*, 7(12):201878, 2020.

[16] Peter Sheridan Dodds and Duncan J. Watts. Universal behavior in a generalized model of contagion. *Phys. Rev. Lett.*, 92:218701, 5 2004.

[17] Peter Sheridan Dodds and Duncan J Watts. A generalized model of social and biological contagion. *Journal of theoretical biology*, 232(4):587–604, 2005.

[18] Joshua M Epstein and Robert Axtell. *Growing artificial societies: social science from the bottom up*. Brookings Institution Press, 1996.

[19] Paul Erdos and Alfréd Rényi. On the evolution of random graphs. *Publ. Math. Inst. Hung. Acad. Sci*, 5(1):17–60, 1960.

[20] Sharad Goel, Ashton Anderson, Jake Hofman, and Duncan J Watts. The structural virality of online diffusion. *Management Science*, 62(1):180–196, 2016.

[21] Jacob Goldenberg, Barak Libai, and Eitan Muller. Talk of the network: A complex systems look at the underlying process of word-of-mouth. *Marketing letters*, 12(3):211–223, 2001.

[22] Mark Granovetter. Threshold models of collective behavior. *American journal of sociology*, 83(6):1420–1443, 1978.

[23] Jon Herman and Will Usher. SALib: An open-source python library for sensitivity analysis. *The Journal of Open Source Software*, 2(9), 1 2017.

[24] Somya Jain and Adwitiya Sinha. Identification of influential users on twitter: A novel weighted correlated influence measure for covid-19. *Chaos, solitons & fractals*, 139:110037, 2020.

[25] Chathura Jayalath, Chathika Gunaratne, William Rand, Chathurani Seneviratne, and Ivan Garibay. A generalization of threshold-based and probability-based models of information diffusion. *Advances in Complex Systems*, 26(2):2350005, 2023.

[26] Chunxiao Jiang, Yan Chen, and KJ Ray Liu. Evolutionary dynamics of information diffusion over social networks. *IEEE Transactions on Signal Processing*, 62(17):4573–4586, 2014.

[27] Maksim Kitsak, Lazaros K Gallos, Shlomo Havlin, Fredrik Liljeros, Lev Muchnik, H Eugene Stanley, and Hernán A Makse. Identification of influential spreaders in complex networks. *Nature physics*, 6(11):888–893, 2010.

[28] Ju Sung Lee, Tatiana Filatova, Arika Ligmann-Zielinska, Behrooz Hassani-Mahmooei, Forrest Stonedahl, Iris Lorscheid, Alexey Voinov, Gary Polhill, Zhanli Sun, and Dawn Cassandra

Parker. The complexities of agent-based modeling output analysis. *The journal of artificial societies and social simulation*, 18(4), 2015.

[29] Pei Li, Jeffrey Xu Yu, Hongyan Liu, Jun He, and Xiaoyong Du. Ranking individuals and groups by influence propagation. In *Advances in Knowledge Discovery and Data Mining: 15th Pacific-Asia Conference, PAKDD 2011, Shenzhen, China, May 24-27, 2011, Proceedings, Part II 15*, pages 407–419. Springer, 2011.

[30] A Ligmann-Zielinska, D B Kramer, Spence Spence Cheruvelil, and K A Soranno. Using Uncertainty and Sensitivity Analyses in Socioecological Agent-Based Models to Improve Their Analytical Performance and Policy Relevance. *PLoS ONE*, 9(10):109779, 2014.

[31] Lei Liu, Ya-Chen Tina Shih, Robert L Strawderman, Daowen Zhang, Bankole A Johnson, and Haitao Chai. Statistical analysis of zero-inflated nonnegative continuous data. *Statistical Science*, 34(2):253–279, 2019.

[32] Jingbo Meng, Wei Peng, Pang-Ning Tan, Wuyu Liu, Ying Cheng, and Arram Bae. Diffusion size and structural virality: The effects of message and network features on spreading health information on twitter. *Computers in Human Behavior*, 89:111–120, 2018.

[33] Giovanni Pegoretti, Francesco Rentocchini, and Giuseppe Vittucci Marzetti. An agent-based model of innovation diffusion: network structure and coexistence under different information regimes. *Journal of economic interaction and coordination*, 7(2):145–165, 2012.

[34] Márton Pósfai and Albert-László Barabási. *Network science*. Citeseer, 2016.

[35] Jiezhong Qiu, Jian Tang, Hao Ma, Yuxiao Dong, Kuansan Wang, and Jie Tang. Deepinf: Social influence prediction with deep learning. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 2110–2119, 2018.

[36] J. O. Ramsay and B. W. Silverman. *Functional Data Analysis*. New York, Springer, 2nd edition, 2005.

[37] William Rand, Jeffrey Herrmann, Brandon Schein, and Neža Vodopivec. An agent-based model of urgent diffusion in social media. *Journal of Artificial Societies and Social Simulation*, 18(2):1, 2015.

[38] William Rand and Roland T. Rust. Agent-based modeling in marketing: Guidelines for rigor. *International Journal of Research in Marketing*, 28(3):181–193, 9 2011.

[39] Andrea Saltelli, Paola Annoni, Ivano Azzini, Francesca Campolongo, Marco Ratto, and Stefano Tarantola. Variance based sensitivity analysis of model output. design and estimator for the total sensitivity index. *Computer physics communications*, 181(2):259–270, 2010.

[40] Susan M Sanchez and Thomas W Lucas. Exploring the world of agent-based simulations: Simple models, complex analyses. In *Proceedings of the Winter Simulation Conference*, volume 1, pages 116–126. IEEE, 2002.

[41] Thomas Schreiber. Measuring information transfer. *Physical review letters*, 85(2):461, 2000.

[42] Chathurani Senevirathna, Chathika Gunaratne, William Rand, Chathura Jayalath, and Ivan Garibay. Influence cascades: Entropy-based characterization of behavioral influence patterns in social media. *Entropy*, 23(2):160, 2021.

[43] Claude Elwood Shannon. A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423, 1948.

[44] Ilya M Sobol. Global sensitivity indices for nonlinear mathematical models and their monte carlo estimates. *Mathematics and computers in simulation*, 55(1-3):271–280, 2001.

[45] Il'ya Meerovich Sobol'. Sensitivity estimates for nonlinear mathematical models. *Matematicheskoe modelirovanie*, 2(1):112–118, 1990.

[46] Forrest Stonedahl, William Rand, and Uri Wilensky. Evolving viral marketing strategies. In *Proceedings of the 12th annual conference on Genetic and evolutionary computation*, pages 1195–1202, 2010.

[47] Steven H Strogatz. *Nonlinear dynamics and chaos: with applications to physics, biology, chemistry, and engineering*. CRC press, 2018.

[48] Beiming Sun and Vincent TY Ng. Identifying influential users by their postings in social networks. In *Proceedings of the 3rd international workshop on Modeling social media*, pages 1–8, 2012.

[49] Jeffrey Travers and Stanley Milgram. The small world problem. *Phychology Today*, 1(1):61–67, 1967.

[50] Greg Ver Steeg and Aram Galstyan. Information-theoretic measures of influence based on content dynamics. In *Proceedings of the sixth ACM international conference on Web search and data mining*, pages 3–12, 2013.

[51] Duncan J Watts. A simple model of global cascades on random networks. *Proceedings of the National Academy of Sciences*, 99(9):5766–5771, 2002.

[52] Duncan J Watts and Steven H Strogatz. Collective dynamics of 'small-world'networks. *nature*, 393(6684):440, 1998.

[53] Rachel Winter, Steve Scheinert, Mel Stanfill, Anastasia Salter, Olivia B Newton, Jihye Song, Stephen Fiore, William Rand, and Ivan Garibay. A taxonomy of user actions on social networking sites. In *Proceedings of the 31st ACM Conference on Hypertext and Social Media*, pages 233–234, 2020.

[54] Simon N Wood. mgcv: Gams and generalized ridge regression for r. *R news*, 1(2):20–25, 2001.

[55] Simon N Wood. *Generalized additive models: an introduction with R*. chapman and hall/CRC, 2017.

[56] S.N. Wood, N., Pya, and B. S"afken. Smoothing parameter and model selection for general smooth models (with discussion). *Journal of the American Statistical Association*, 111:1548–1575, 2016.

[57] Luo Xiao, Vadim Zipunnikov, David Ruppert, and Ciprian Crainiceanu. Fast covariance estimation for high-dimensional functional data. *Statistics and computing*, 26:409–421, 2016.