Tracking Infection Diffusion in Social Networks

by

Tavis Joseph Pedersen

BASc. Engineering Physics, University of British Columbia, Canada, 2015

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF
THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF APPLIED SCIENCE

in

The Faculty of Graduate and Postdoctoral Studies

(Electrical and Computer Engineering)

THE UNIVERSITY OF BRITISH COLUMBIA

(Vancouver)

August 2017

Abstract

This thesis explores the problem of tracking the diffusion of contagion processes on social networks. Infection (or Information) diffusion is modeled using the Susceptible-Infected-Susceptible (SIS) model. Mean field approximation is employed to approximate the discrete valued infection dynamics by a deterministic ordinary differential equation, thereby yielding a generative model for the infection diffusion. The infection is shown to follow polynomial dynamics and is estimated using an exact non-linear Bayesian filter. We compute posterior Cramér-Rao bounds to obtain the fundamental limits of the filter which depend on the structure of the network. The SIS model is extended to include homophily, and filtering on these networks using the exact non-linear Bayesian filter is illustrated. With consideration for the collaborative or antagonistic nature of some social processes on networks, we present an alternative, game theoretic, model for the spread of information based on the evolutionary Moran process. The diffusion of collaboration following this Moran model is estimated using a particle filter. Additionally, we validate the efficacy of our method with diffusion data from a real-world online social media platform, Twitter. We find that SIS model is a good fit for the information diffusion and the non-linear filter effectively tracks the information diffusion.

Lay Summary

In this thesis we use expert knowledge in the form of models for infectious disease in combination with statistical algorithms to better understand how an infection is spreading and who the afflicted are. Electrical engineers are practiced at using models of underlying systems to combine observations and form predictions, and these techniques can be extended beyond the world of electrons, bits and binary. By better understanding the flow of viruses and disease we can better mobilize our knowledge and efforts to manage, slow and eventually eradicate microscopic dangers. While these contagions pose real threat, they are not as ever-present as other viral phenomena in the developed world. Our lives are more immediately impacted by viral news, entertainment, and opinion. The same methods which allow us to better understand the presence of Ebola in a population can be utilized to better understand the population's outlook on, or exposure to, a political movement.

Preface

The following publications have resulted from the research presented in this thesis:

- V. Krishnamurthy, S. Bhatt, and T. Pedersen, "Tracking Infection Diffusion in Social Networks: Filtering Algorithms and Threshold Bounds", *IEEE Transactionson Signal and Information Processing over Networks*, vol. 3, no. 2, 2017

Statement of Authorship

I am a co-author of the publication listed above. I have been responsible for developing original ideas, deriving mathematical solutions, and generating simulation results for this publication with valuable frequent suggestions, advice, contributions, and feedback from Prof. Vikram Krishnamurthy and Sujay Bhatt.

A version of the content presented in the Chapter 4 has been submitted as a course project for *UBC MATH 564: Evolutionary Dynamics.*

Table of Contents

Appendix

List of Tables

List of Figures

List of Abbreviations

SIS	:	Susceptible-Infected-Susceptible
MFT	:	Mean Field Theory
MFD	:	Mean Field Dynamics
CRLB	:	Cramèr-Rao Lower Bound
PCRLB	:	Posterior Cramèr-Rao Lower Bound
MCMC	:	Markov Chain Monte Carlo
RDS	:	Respondent Driven Sampling
HMM	:	Hidden Markov Model
ER	:	Erdős-Rényi
SF	:	Scale-Free
MSE	:	Mean Square Error

Acknowledgements

Firstly, I would like to thank my advisor, Professor Vikram Krishnamurthy, whose curiosity, pragmatism and dedication to excellence have elevated my own work and continue to elevate the work of those around him. Without the intellectual and financial support of Dr. Krishnamurthy, this work would never have been possible. I also owe considerable thanks to Sujay Bhatt, whose abilities cannot be overstated. Sujay's talented mind, patience, and sense of humour were beacons which guided me during my journey through my MASc.

I would like to thank the professors and administration of the University of British Columbia. My educational experience has been enriched immeasurably by the care and support of this exceptional group. The brilliance of the Faculty under whom I have studied, researched, and worked has unveiled the abundance of joy that can be found in the undying pursuit of knowledge.

I would also like to thank my colleagues in the Statistical Signal Processing laboratory for their support and many kindnesses during my MASc. Furthermore, I am grateful to Anahita Shojaei, Philip Edgecumbe, Malik Nasser, Buddhika Netasinghe, Jorden Heatherington, and Felix Funk for their friendship and academic support during my MASc research. Outside the boundaries of the university, I lend my thanks and love to my many friends who have carried me, taught me, and with whom I have shared so much joy.

Finally, I am unendingly grateful for the love and support of my mother, brothers, and sister. My family has always been the foundation of all my strength, and anything that I accomplish for good in my life will have been because of them.

Dedication

To my cherished mother Cheri Pedersen, brothers Liam and Tyler Pedersen, and sister Camryn Pedersen

With love and deepest appreciation.

Chapter 1

Introduction

1.0.1 From Electrical Engineering to Infectious Disease

Compared to social, political and economic structures, which are often influenced by enigmatic desires or hidden impetuses, the world of electrical signals is relatively straightforward. We have a good and ever improving understanding of the physical processes that underlie a signal's generation and propagation. We can, for instance, look to Schroödinger's equation or the telegrapher's equations to understand electrical phenomena on scales as small as atoms or as large as international electrical grids. We have studied for centuries these phenomena and can summarize their states with words like phase, voltage, current, potential, and field strength. These are all measurable quantities and, if we want to know the voltage of a line on a computer, we can make use of tools such as a digital voltmeter or a galvanometer (an old type of voltmeter). Sometimes we need to make measurements quickly and accurately, more accurately than our tool can provide in any one measurement. We can utilize our knowledge of electrodynamics and multiple measurements to improve our estimate. If we measure the current of a power line directly and measure the magnetic field at some known distance from the line, we know, through Maxwell's Equations, that our measurement of the magnetic field provides information about the current in the line. Thus, we have multiple sources of information about a signal. The combination of imperfect information into better and better estimates through knowledge of the underlying system is the foundation of statistical signal processing. Electrical signals were a natural cradle for the development of these processing algorithms, but the same algorithms that allow us to better estimate the voltage of a connector on a motherboard can be used to combine expert knowledge and imperfect measurements in other areas. In this thesis, we consider measurements of the levels of infection in a population. We can combine models borrowed from epidemiology with signal processing algorithms to eke out more information from any measurements of how an infection is propagating through a population. By doing this we heighten the value of every data point collected. A longitudinal survey of communities in a disease ridden area can be used in combination with model informed algorithms to more accurately estimate how many people are infected and which subpopulations are infected. Furthermore, we realize that epidemiological models are increasingly useful outside of epidemiology. The internet has many viral phenomena, which spread in a manner similar to diseases. These social infections can relate to politics, marketing, religion, or entertainment and play an increasingly large role in our lives as individuals and society as a whole.

1.1 Overview

The dark spectre of disease has stalked humanity since our collective infancy. Mankind's greatest enemy in the kingdom Animalia is neither the noble tiger nor the colossal great white. These beasts, as terrifying as they may have been to our ancestors, pose no more threat to humanity than falling vending machines. Our greatest animal adversary is the petite and easily squished mosquito. It is through disease that the mosquito, a potential vector of many illnesses, earns this distinction. More deadly than the conquistadors and their guns were their germs. More deadly than the trenches and munitions of the First World War were the beautifully and woefully efficient distribution systems that brought men from every corner of the world to the European theater and sent back the Spanish Flu. For most of human history, mankind's relative immobility has been a bulwark against the spread of disease. Even though trade did spread the Black Death throughout Asia and Europe, the plague spread at the speed of sail and could only fly as fast as the feet that carried it. Today, those feet can travel more quickly than ever. A flight attendant might have breakfast in Abu

Dhabi, lunch in Frankfurt, and find themselves in California the following day. This poses a new-dramatic risk for humankind, how can we defend against an infectious disease in a world that is more connected than ever before? Scares such as SARS, the Avian influenza and Ebola have garnered public interest, even panic, but have not yet laid down our defenses. Today, we fight the spread of these diseases and infections through the efforts of medical professionals and epidemiologists. Quarantines, vaccines, massive laboratory efforts all serve to stem the tide of infectious disease, but more tools are necessary in this fight. We need to better understand the spread of infections if we are to effectively curb the next microscopic threat to humanity, which may be frozen just beneath some now precariously warm layer of permafrost. To better understand the spread of infections, we need to track them in real time. We must gather information about who is infected, how quickly the infection is spreading and who we can target with health services to most effectively diminish its spread. The filtering algorithms of engineering literature have a role to play in this titanic effort. Through statistical algorithms, we can more accurately estimate the level of infection in populations. It is the development of this idea that we pursue in this thesis.

The speed of infection spread depends upon the structure of human interactions. As people become more connected, infections can more easily spread between distinct portions of the population. The structure over which human interactions occur is the human social network. Our local social environments dictate our exposure to new information, susceptibility to disease, and the network of these connected local environments determines how vulnerable an entire population is to contagion. In this thesis, we will speak of infection, cooperation, and information diffusion as analogous processes, a parallel motivated by the knowledge that viral phenomena online behave similarly to virulent disease spreading throughout a population [74]. We therefore extend analysis performed on biological networks and epidemic models to the spread of knowledge and technology through social and online networks [58]. Motivating the inclusion of these more generalized contagions is the importance of understanding the spread of technologies, social learning theory, and the economic theory of interacting agents.

Further to this motivation, the statistical filtering techniques of this work are more readily applicable in the context of social contagions than to classical disease propagation[1].

When considering adoption of a new technology, influence often seeps in from those around us, indirectly by observing their behaviour or directly through recommendations. This diffusive tide of information, central to technological adoption, has been studied in [18, 31] where the diffusion of hybrid corn and CNC machines were investigated, respectively. The application of epidemic and, in particular, Susceptible-Infected-Susceptible (SIS) model approximations to the diffusion of technology are considered in [42, 55, 59].

The adoption of a technology or behaviour because of a neighbour's adoption can be viewed as an act of social learning. Adoption is an action which represents a piece of information about the technology or behaviour itself and its quality [18]. The diffusion models considered in this work can therefore be considered in the context of agents who are learning in a social network. The statistical methods presented for tracking infections are thus applicable to tracking the process by which a population learns [49]. These diffusion and learning processes depend upon interaction between and influence by individuals, or agents, in the network. Agent based models have also been used to model infection diffusion [72], highlighting the interrelation between the two models.

Models for infection and information diffusion on social networks utilize established knowledge to predict how these processes spread, however, in the real world, randomness can dominate the ultimate outcome. Consumers may flock to a product for unpredictable reasons, and diseases may be spread or stopped by the whims of a single traveler. Because of these random factors, we cannot formulate models which can predict infection/information diffusion exactly. To better estimate the state of the diffusion process, we utilize observations from the real world. These observations can come in the form of polls of political preferences or surveyor information reporting on the state of an epidemic. Since information can rarely

[1] Health related fields adopt new technologies at a highly conservative pace. Significant further research and development of the ideas presented in this thesis are necessary before the work can be applied in the context of mitigating contagion and saving lives. This work serves as a stepping stone, by which we stride closer to that ultimate goal.

be gathered about the entire population of interest, observations provide noisy (containing error) estimates. To combine the expert information engrained in diffusion models with the crucial updates provided by real world observations, we use tracking algorithms borrowed from previous work in statistical signal processing.

In this thesis, we consider the problem of attempting to track the diffusion of information or infections on social networks. We formulate a model for infection diffusion, apply Bayesian filtering algorithms to estimate the infected population state, provide extensions to the infection diffusion model, and finally demonstrate the filtering methods on a dataset derived from analyzing real world tweets relating to the 2011 Arab Spring Uprising.

1.1.1 SIS model and Mean Field Dynamics

Individuals interacting with one another may share knowledge, pathogens, political opinions, or particularly funny jokes. Common to all of these is an underlying social mechanism of transfer: Contact is made, either online or in person, and the phenomena spreads. By formulating and analyzing a mathematical framework for contact diffusion processes, we may reveal a realm in which we can better understand and control their spread. A natural starting point is in epidemiology, the study of the spread of infectious disease. Infectious diseases have a profound impact on human societies. Some infections insidiously lay within a population and damage their hosts without warranting action on a massive scale, while others collapse towns, cities and even civilizations. In the threat they pose, they gather a great deal of interest. One attempt to model infectious disease is the SIS model [71]. In the SIS model, an individual who is susceptible may become infected if they come into contact with another infected individual and, should they become infected, may recover to become susceptible again. The transitions from infected to susceptible (and vice versa) are assumed to be probabilistic. In the SIS model, real world details about specific interactions between individuals, for instance whether an individual sneezed during a encounter, are encapsulated into a single transition probability. The scenario outlined by the SIS model emulates a disease which may be inflicted upon the same individual repeatedly - superficially, like the common cold. The evolution of such an infection occurs over a structured population. This structure is formed of social connections. Our friends, family, acquaintances, and colleagues form our social environment, and the overlap of all social environments form a social network. As described in [59], there is a wide range of social phenomena such as diffusion of technological innovations, cultural fads, and economic conventions [18] where individual decisions are influenced by the decisions of others. A large body of research on social networks has been devoted to the diffusion of information (e.g., ideas, behaviors, trends) [34], and particularly on finding a set of target nodes so as to maximize the spread of a given product [19, 65].

In the SIS model, given an individual, its neighbourhood and its current state, its next state can be considered a weighted coin toss. While these individual interactions are extremely simple, when the entire network is considered, the exact spread of the infection is combinatorially complex. In statistical physics, a similar scaling of complexity was encountered in the study of spin glasses, lattices of particles with interacting magnetic dipoles. Except in trivial structures, such as lines, the spin systems could not be analyzed exactly. To model these systems, physicists replaced all the exact interactions between particles with interactions between individual particles and the ensemble average of all other particles. This approximation is called mean field theory (MFT) and can be applied to other many particle systems, such as the SIS model.

We apply a mean field theory that discriminates between nodes of differing degree. Through this discrimination, we are able to encode network structure into the dynamics of the infection diffusion. These degree-based mean field dynamics (MFD) approximations for SIS model have been derived in [59, 74].

There are several models for studying the spread of infection and technology in complex networks including Susceptible-Infected-Susceptible (SIS), Susceptible-Alert-Infected-Susceptible (SAIS), and Susceptible-Exposed-Infected-Vigilant (SEIV); see [27, 41]. Susceptible-Infected-Susceptible (SIS) models have been extensively studied in [42, 49, 58, 59, 71, 82, 83, 88] to model information/infection diffusion, for example, the adoption of a new technology in a consumer market. In this thesis we follow a similar approach and utilize mean field dynamics to approximate the discrete-valued infection dynamics by a deterministic ordinary differential equation. The mean field dynamics yield a tractable model for Bayesian filtering in order to estimate the infection dynamics given a sampling procedure for the social network.

1.1.2 Non-linear filter and PCRLB for Bayesian Tracking of Infected Populations

The infection (or information diffusion) dynamics of the SIS model, and the underlying diffusion processes which it represents, are non-deterministic. Every disease, product launch, rumor, or news piece is influenced by factors which can best be described as random. These random factors make it impossible to predict apriori (before observation) how many individuals in a network are infected. Since even precise models would be unable to predict these random elements, we are motivated to observe the state of the system directly. These observations may be formed by sampling [2] a representative subset of the population. If every individual in the network could be sampled, we would know the precise infected population state at that instant. Unfortunately, sampling can be expensive or prohibitively difficult. In the case of infection, sampling may require physically interacting with infected individuals; a dangerous task for those involved. Additionally, the individuals we want to sample may not be known to us. For example, we may be interested in learning about how infection spreads amongst drug users. In this case, members may be willing to anonymously share information for compensation, but the structure and members of this population may be unknown to researchers. For all of these cases, sampling an entire population at one time instant may not be possible, and repeatedly sampling all individuals to track an infection over time is even

[2] We use sampling to refer to learning the state of individuals by some means. The mechanism of learning is not specified but may be surveying or querying directly.

more challenging. The SIS model for the underling infection dynamics, discussed in Sec. 2.2, contributes information about how the infection evolves. We can utilize this information to predict how the state evolves. To combine this model based knowledge of state evolution with the direct observations of the state, we can use a Bayesian filtering algorithm.

Bayesian filtering algorithms allow us to combine observations with apriori state estimates to form a posteriori (after observation) estimate. The most commonly known Bayesian filter is the Kalman filter. The Kalman filter is a finite dimensional optimal filter that combines noisy (contains random errors) linear observations of a linearly evolving noisy (affected by random changes) state. It assumes Gaussian observation noise and Gaussian state noise. When the state evolution is not linear, or noise is non-Gaussian, other filtering algorithms may produce more accurate estimates of the state. The SIS dynamics described in Sec. 2.2 are not linear, but we show that the Mean Field Dynamics (MFD) are polynomial in the state. That is, in MFD the state at timestep $n + 1$ depends polynomially on the state at time n. We utilize a Bayesian filter which computes an optimal estimate of the state in polynomially evolving systems with Gaussian state and observation noise.

Now that we have a filtering algorithm, we are interested in knowing how “good” the estimates generated by these filtering algorithms are. We compute the Posterior Cramer-Rao Lower Bound (PCRLB) of the non-linear filter. The PCRLB provides a lower bound on the error covariance of our filter estimate. We examine via numerical examples and posterior Cramér-Rao lower bounds, how state estimation over large networks is affected by the network; see [79] for posterior Cramér-Rao bounds for non-linear filters. Numerical examples illustrate the difference in performance between power law (scale-free) and Erdős-Rényi graphs.

In this thesis, since we focus on optimal Bayesian estimation (filtering) of the infection dynamics by sampling the network, we use the SIS model. In small sized sensor networks represented by graphs, [75] derived an optimal estimation algorithm, based on the Kalman filter, to estimate the state at each sensor. Further, on specific structures like trees, [75] provides expressions for the steady-state covariance; see also [69]. Finally, for Kalman filtering of the Susceptible-Exposed-Infected-Recovered model, see [26].

1.1.3 Extended Models in Network Analysis: Homophily and Game Theoretic Dynamics

The infection diffusion of Chapter 2 models contagions which spread only as a function of the immediate local environment of each individual. This model is sensible for diseases and some types of information flow, however, it can fail to represent interactions between self-interested, rational characters who are interacting on a cooperative or competitive way. Consider the scenario where your decision depends upon the decisions of your neighbours, and you neighbour's decision on their neighbours, in this scenario each agent's actions depends not only on their immediate environment, but also on the network as a whole. Take for instance, the adoption of a new smart phone, the strategy of an individual [buy, not buy] may be dependent on their neighbourhood. Do they have a significant number of friends with the new phone with whom they can take advantage of shared features, or is the quality (or fitness) of such a strategy insufficient to warrant the new adoption? They may choose to copy the choice of their friends, but they may also take into account how happy their friends are with their current phones, which in turn depends on the current state of all their neighbours. In this way, the fitness of any one individual's choice is effected by the entire state of the network.

Game theory provides a mathematical framework for interactions between rational self interested persons, where each involved individual, or player, attempts to predict the choices of its counterpart and act accordingly to maximize its own reward. Evolutionary game theory was developed as a framework for researchers to explore the survival of traits or strategies within a population in a biological context. In evolutionary game theory, individuals in a population have traits or strategies which my promote, suppress or be uninfluential in the individual's survival, ultimately affecting their survival fitness in the population. Importantly, evolutionary graph theory explores the temporal trend of these populations as individuals of each type die and reproduce at a rate which is related to their fitness. Concepts such as fitness, birth and death are easily generalized to non-biological systems, where fitness may be related to monetary/wellbeing outcomes and birth and death may be of ideas, rather than beings. Extending the example of consumers deciding whether or not to choose a smartphone, we may consider the level of satisfaction their friends have with their current smartphones. In this scenario, satisfaction is analogous to biological fitness. An individual is more likely to copy the smartphone choice, or strategy, of friends who they know to be highly satisfied with their smartphones.

The model of infection diffusion considered primarily in this thesis was the SIS model, however, these models do not consider the fitnesses of individuals and their state. To incorporate this fitness information into interactions we consider the Moran process [73]. The fitness of individuals is modeled by the predicted payoff of a two player game.

In this thesis, we consider the role of network structure in both the pair approximation as well as the mean field approximation. We develop mean field and pair approximation models for the Moran Process. Mean field models are developed for death-Birth updating, Birth-death updating and for death-Rebirth updating (Diffusion). A heterogeneous model is also developed for death-Birth updating in the Pair Approximation. All of these frameworks then admit the investigation of the effects of network structure on the long term levels of cooperation. The deterministic trajectories of these models are then compared against simulated trajectories of the level of cooperation in a synthetic network with the desired degree distribution. Further, we highlight that the mean field and pair-approximation methods are not amenable to the optimal filtering of Chapter 3, but the infected population state (or cooperation state) can be estimated using particle filtering methods.

1.1.4 Numerical Results

In Sec. 2.2 of this thesis, we provide theoretical justification for using a mean field dynamical model of the state evolution, and in Sec. 3.1 we utilize this model in a filtering algorithm. The theoretical justification relies upon the assumptions of an accurate model for the underlying system, sufficiently large network size, sufficiently long time horizon, and Gaussian state and observation noise. To explore the validity of these assumptions we illustrate the performance of the filtering algorithm on synthetic and real world data. Sec. 5.2 illustrates the SIS model and the performance of the Bayesian filter on simulated data and examines the sensitivity of the filter to the underlying graph model (Erdős-Rényi vs scale-free). Sec. 5.4 also presents a detailed example using a real Twitter dataset. It is shown via a goodness of fit test that SIS is a reasonable model for information propagation in the Twitter dataset and that the infected population state can be tracked satisfactorily over time via the Bayesian filter.

1.2 Main Contributions

In the following subsections, a brief list of contributions of each chapter of this thesis is provided. Detailed discussions of each contribution are provided in the corresponding chapters.

1.2.1 SIS model and Mean Field Dynamics

As explained in Section 1.1.1, in Chapter 2, a formulation for infection diffusion based on the SIS model is developed, and the mean field dynamics of the model are presented. The main contributions of this chapter are:

1. A formulation of discrete time infection dynamics on a social network based on the SIS model.

2. Showing that mean field dynamics can be used to approximate the discrete time infection dynamics with a deterministic differential equation.

3. Presenting two sampling methods that can be used to observe the infected population state of a social network.

1.2.2 Non-linear filter and PCRLB for Bayesian Tracking of Infected Populations

As explained in Section 1.1.2, in Chapter 3, a non-linear Bayesian filter is presented and applied to the infection diffusion model of 2. The main contributions of this chapter are:

1. A formulation of the non-linear Bayesian filter applied to an infection diffusion model
2. The Posterior Cramer Rao Lower Bound of the Bayesian filter is computed for synthetic networks and compared to the mean square error of the filter estimate for both scale-free and Erdös-Renyi networks.

1.2.3 Extended Models in Network Analysis: Homophily and Game Theoretic Dynamics

As explained in Section 1.1.3, in Chapter 4, we consider extensions and alternatives to the SIS model of Chapter 2. The main contributions of this chapter are:

1. An extension to the SIS model to include network homophily is presented.
2. A framework based upon the Moran process in evolutionary game theory is presented for the spread of infection.
3. The mean field and pair approximations to the Moran process are extended to complex networks.
4. A particle filtering method is presented for filtering game theoretic network dynamics.

1.2.4 Numerical Results

As explained in Section 1.1.4, Chapter 5, the filter developed in Chapter 3 is applied to simulated and real world infections. The main contributions of this chapter are:

1. The effect of model misspecification and performance of the filter discussed in Chapter 3 is examined.
2. Numerical examples of the model extensions presented in Chapter 4 are presented.
3. The performance of the filter discussed in Chapter 3 when applied to a real world Twitter dataset is presented.

1.3 Outline of the Thesis

The outline of the rest of this thesis is as follows.

Chapter 2 formulates a model for representing the diffusion of infection, information, or opinion on a social network. First, we provide a mathematical description of a social network and introduce two specific types of networks: Erdös-Renyi and scale-free. Next we present mean field theory for interacting particle systems. Then, we present a formulation of infection diffusion model based on the SIS model. We show that this discrete time model can be approximated by an ordinary differential equation using mean field theory, and that the mean field system evolves with polynomial dynamics. Finally, two sampling methods are presented for observing the infected population.

Chapter 3 considers the problem of attempting to track the infection diffusion modeled in Chapter 2. It begins by formulating the mean field dynamics as a polynomial state space evolution with observations given by the sampling methods of Chapter 2. An optimal non-linear Bayesian filter is then presented for estimating the infected population state. The PCRLB is calculated to give a theoretical performance limit of the filtering algorithm for Erdös-Renyi and scale-free network types.

Chapter 4 considers an extension to the infection model presented in Chapter 2 to model networks with homophily, as well as an alternative game theoretic model based on work done in Evolutionary Dynamics. The evolutionary model allows for the representation of competition within the population, expanding the scope of diffusive scenarios which can be accurately approximated. The problem of tracking the infected population state is considered for both models; limitations of the Bayesian filter are considered for the game theoretic model and an particle filter is presented to track the infected population state.

Chapter 5 presents numerical simulations and investigations to establish the performance of the filters described in this thesis. First effect of misspecified models and sampling on filter performance is investigated. Further, numerical examples of filtering on networks with homophily and on networks with game theoretic diffusion are presented. Finally, a real world dataset from Twitter is used to demonstrate the performance of the filter.

Finally, Chapter 6 provides a conclusion of the work presented in this thesis and some future research directions.

Chapter 2

SIS Model and Mean-field Approximation

This section discusses the discrete time SIS model [4], [17] and mean field dynamics for the diffusion of information in a social network.

2.1 Social Networks and SIS dyamics

2.1.1 Social Networks

Humans are social creatures, and prior to the information age, social networks were predominantly the networks formed by connection in the physical world. In this social network, an individual might be considered connected to another if the two know each other, met regularly, or exchanged letters. With the rise of the Internet, the term “Social Networks” has became ubiquitous. It, or its translated variants, are used around the World, and massive Online Social Networks have become a cornerstone of communication, entertainment, information dissemination, and, increasingly, politics. In 2017, among the most dominant Social Networks are Internet leviathans Facebook, Twitter, RenRen, Sina Weibo, and Linkedin. The usage of online Social Networks span socio-economic condition, class, education, and interest. The ubiquity of social networks motivates increased analysis to comprehend how these networks effect our lives on a personal and societal level. The processes that occur on Social networks are often dependent on the network structure, and so understanding the interplay between the social network structure and the processes that evolve on the social network is critical to understanding the spread of disease, political influence, and technological adoption.

Mathematically, a social network is represented by graph G:

$$G = (V, E), \text{ where } V = \{1, \dots, M\}, \text{ and } E \subseteq V \times V. \tag{2.1}$$

Here V denotes the set of M vertices (nodes) and E denotes the set of edges (relationships). The degree of a node m is its number of neighbors[3]:

$$\Xi^{(m)} = |\{g \in V : g, m \in E\}|, \ | \cdot | \text{ denotes cardinality.}$$

$M(d)$ denotes the number of nodes in the social network G with degree d, and $\rho(d)$ denotes the degree distribution. That is, for $d = 0, 1, \dots, D$,

$$M(d) = \sum_{m \in V} \mathcal{I}\big(\Xi^{(m)} = d\big), \quad \rho(d) = \frac{M(d)}{M}.$$

Here $\mathcal{I}(\cdot)$ denotes the indicator function and D denotes the maximum degree. Since $\sum_d \rho(d) = 1$, $\rho(d)$ can be viewed as the probability that a node selected randomly on V has degree d. The degree distribution of a network is a powerful summary measure which gives insight into connectedness and structure of networks. Throughout this work, we concern ourselves with complex networks. A complex network is a network with non-trivial network topological features, such as a high degree of clustering or heavy-tailed degree distribution. In particular, in this thesis, we consider complex networks with two commonly studied degree distributions: Erdös-Renyi and scale-free networks.

[3] A vertex g is adjacent to a vertex m if there is an edge between them. The neighbors of a node m are all the vertices that are adjacent to m.

Erdös-Renyi Networks

Erdös-Renyi networks are generated by adding links between nodes randomly with some probability p. These networks are predominantly of interest because of how well studied they are in literature. The degree distribution of Erdös-Renyi networks is

$$\rho(d) = \binom{M-1}{d} p^d (1-p)^{M-1-d} \tag{2.2}$$

Scale-Free Networks

A scale-free network is a network with a degree distribution that follows a power law, that is, $\rho(d) \propto d^{-\gamma}$. These networks are of interest because many real world networks are know to be scale-free. Notably, network of outgoing links on the Internet has a scale-free degree distribution [8], and an example considered in this work, the Social Network Twitter has a scale-free Degree distribution.

2.1.2 Infection Processes on Social Networks: SIS Model

We desire to know how infectious processes evolve on Social networks. Take for instance the Ebola epidemic that occurred in West Africa from 2013-2016. Ebola is a virulent disease that spreads when an uninfected person comes into contact with the bodily fluids of an infected individual. Infection can then only spread by people who have a physical connection: their neighbours in a network. Understanding this contact process and the role of network structure in the process dynamics may give researchers and health workers insight into how to slow, stop, and prevent the spread of such a deadly virus. Modeling such a process is a fundamental problem in epidemiology. One simple, but heavily investigated model is the Susceptible-Infected-Susceptible (SIS) model [4] [17]. In the SIS model, individuals can be infected or susceptible, that is, they are in a state which is susceptible to becoming infected.

At each time instant $n = 0, 1, \ldots$, an individual, or node, m has states

$$s_n^{(m)} \in \{1 \text{ (infected) } 2 \text{ (susceptible) } \}.$$

and the entire population can be described by the *infected population state* at time n, denoted by $\bar{x}_n$, the fraction of nodes with degree d at time n that are infected. That is[4]

$$\bar{x}_n(d) = \frac{1}{M} \sum_m \mathcal{I}\big(\Xi^{(m)} = d, s_n^{(m)} = 1\big), \quad d = 1, \ldots, D. \tag{2.3}$$

Infection can pass from infected individuals to their susceptible neighbours. In this thesis, this infection process is modeled probabilistically. At any discrete timepoint a susceptible individual, or node, has some probability of becoming infected, dependent on the number of infected neighbours and total neighbours it has. An entire population can then be described as a state of a Markov chain where the transition probabilities depend on the degree of the node and the number of infected neighbors.

The infection dynamics evolves as follows[5,6]: :

Step a.) A node m is chosen uniformly from the vertices in V. Suppose the node has degree $\Xi^{(m)} = d$ and the number of infected neighbors $F_n^{(m)} = a$. The state $s_n^{(m)}$ of the node m at time n evolves from state i to j with transition probabilities $\bar{P}_{ij}$, where $i, j \in \{1, 2\}$. These transition probabilities depend on the node's degree d and its number of infected

[4] Note that $M(d) = \rho(d)M$ and the infected population state is scaled by $\rho(d)$ to have a common denominator.

[5] The state of the network (2.1) at time n is given by an M dimensional vector with elements 1 or 2. The state space is given by $\{1, 2\}^M$ and the dynamics can be formulated in terms of a transition matrix between all possible states [74],[78]. In this thesis, we group the 2^M states into D subsets, one for each degree, resulting in a state space of dimension $2^{M(d)}$ for each degree d.

[6] As an example, suppose there are $M = 100$ nodes, with 30 nodes having degree 1 and 70 nodes having degree 2. Then $d \in \{1, 2\}$, $M(1) = 30$, $M(2) = 70$. Suppose the fraction of infected nodes of degree 1 and 2, are respectively, $\bar{x}_n(1) = 3/30$, $\bar{x}_n(2) = 60/70$. Then $\bar{x}_{n+1}(1) \in \{2/30, 3/30, 4/30\}$ if a node of degree 1 is chosen in Step (a), while $\bar{x}_{n+1}(2) \in \{59/70, 60/70, 61/70\}$ if a node of degree 2 is chosen in Step (a).

neighbors a as

$$\bar{P}_{ij}(d,a) = \mathbb{P}\left(s_{n+1}^{(m)} = j | s_n^{(m)} = i, \Xi^{(m)} = d, F_n^{(m)} = a\right). \tag{2.4}$$

In real world applications, these transition probabilities can be chosen by experts, such as a team of virologists who can estimate the probability that an individual interacting with an Ebola infected person becomes, themselves, infected, or determined empirically, if many individual interactions can be observed.

Step b.) The population state of degree d is updated as

$$\begin{bmatrix} \bar{x}_{n+1}(d) \\ 1-\bar{x}_{n+1}(d) \end{bmatrix} = \begin{bmatrix} \bar{x}_{n}(d) \\ 1-\bar{x}_{n}(d) \end{bmatrix} + \frac{1}{M} \times \begin{bmatrix} \mathcal{I}\big(s_{n+1}^{(m)} = 1, s_n^{(m)} = 2\big) - \mathcal{I}\big(s_{n+1}^{(m)} = 2, s_n^{(m)} = 1\big) \\ \mathcal{I}\big(s_{n+1}^{(m)} = 2, s_n^{(m)} = 1\big) - \mathcal{I}\big(s_{n+1}^{(m)} = 1, s_n^{(m)} = 2\big) \end{bmatrix}. \tag{2.5}$$

Here $1-\bar{x}_{n+1}(d)$ denotes the fraction of susceptible nodes of degree d at time $n+1$. According to (2.5), the infected population state, $\bar{x}_{n+1}$, changes by $\frac{1}{M}$ at every time-step depending on the transition probabilities.

Here we have presented the SIS model with respect to its classical application in infection, but we can similarly model the spread of influence, interest, or ideas online. Information can be passed between individuals, for example by web posts or tweets, and this information spreads in a way analogous to an infection. A user may tweet about a news event, for example a shark attack in Sweden. This would be sensational, and the contacts of this user may see that tweet and share or produce their own tweets about the topic. These people may then continue to share, and so on and so forth, as the infection propagates throughout the online social network. This process is known to spread in a way that mimics viral infections [84].

2.1.3 The Mean Field Approximation

The SIS model provides a framework for understanding the spread of an infection, but in practice, even for small network size, there are too many interactions to account for exactly, and the dimensionality of the SIS model, namely 2^M- states, becomes intractable for modeling or signal processing algorithms. The social networks we would like to consider are sometimes massive, such as the Facebook Social network of 1.86 billion active users [1]. To hope to utilize this model in any real world applications, we must use an approximation. For this we look towards mean field theory, and in this section we present a mean field dynamics approximation to the SIS model.

Mean field theory was developed in statistical physics to approximate interactions between large number of particles. Instead of modeling interactions between individual particles, mean field approximations simplify the process by having particles interact with the ensemble averages of all other particles. In social network analysis, the we apply this technique to groups of individuals, instead of particles. We consider the propagation of a process spreading throughout the network, for instance a virulent disease or infection.

In its most simplistic form, in the mean field evolution of a system, a summarizing statistic of the network, denoted here by α, evolves as a function of only itself:

$$\alpha_{n+1} - \alpha_n = \theta \left[\mathcal{T}^+(\alpha_n) - \mathcal{T}^-(\alpha_n) \right] \tag{2.6}$$

Here $\mathcal{T}^+$ is the probability that α_n increases by θ, and $\mathcal{T}^-$ is the probability that α_n decreases by θ. $\theta > 0$ is the diffusion parameter which corresponds to the fraction of the population that interacts per timestep. This diffusion parameter, along with the transmission probabilities define the timescale of the infection[7].

This simple mean field theory does not permit a description of network structure beyond

[7] Here we further note that θ may be dependent on time, n, and thus may non-linearly rescale the evolution of the system. If, for instance, interactions between individuals occur more frequently on a Saturday than a Tuesday, this information can be encoded into θ, to allow for more complex modeling.

regular networks and thus may fail to capture structures that may change the rate of of infection. For example, these mean field models do not adequately account for influential individuals who are well connected in a network. Consider a Twitter user with a great number of followers, if they adopt an idea, they may dramatically increase the rate at which that idea spreads. Similarly, a social individual may singlehandedly sustain the spread of an infection by repeatedly interacting with a great number of susceptible persons. To encode some of these network effects, we consider a more complex mean field description when modeling infection on networks.

Mean Field Approximation with Network Structure

Mean field theory can incorporate information about the network degree distribution to allow for the modeling of complex networks [86]. To represent a simple mean field system we utilized a single summarizing statistic, α. We can incorporate network structure into the mean field dynamics by formulating the evolution of a vector of summarizing statistics. In this case we will consider the infected population state, which is differentiated by degree $\bar{x}_n(d)$. Now the change in each infection probability $\bar{x}_n(d)$ depends on degree d and the entire current state $\bar{x}_n$:

$$\bar{x}_{n+1}(d) - \bar{x}_n(d) = \theta \left[\mathcal{T}^+(\bar{x}_n, d) - \mathcal{T}^-(\bar{x}_n, d) \right] \tag{2.7}$$

Here θ is the fractional change per interaction, and $\mathcal{T}^+(\bar{x}_n, d)$ and $\mathcal{T}^+(\bar{x}_n, d)$ denote the probabilities that the fraction of infected nodes of degree d increases or decreases, respectively. This incorporation of degree information allows us to model networks and differentiate between networks based on their degree distributions.

2.2 SIS Model and Mean Field Dynamics

If we again consider our social network, over which an infection evolves according to SIS dynamics, we can express mean field dynamical equation in terms of our infected population states $\bar{x}_n$.

The probability that a link, chosen uniformly at random among all links of a node chosen uniformly at random in the network at time n, points to an infected node is given by $\alpha(\bar{x}_n)$ as the *infected link probability.* Clearly,

$$\begin{aligned}\alpha(\bar{x}_n) &= \frac{\sum_{d=1}^{D} (\text{\# of links from infected node of degree } d)}{\sum_{d=1}^{D} (\text{\# of links of degree } d)} \\ &= \frac{\sum_{d=1}^{D} d\,\rho(d)\,\bar{x}_n(d)}{\sum_{d=1}^{D} d\,\rho(d)}.\end{aligned} \tag{2.8}$$

With α_n, the probability that a susceptible node with d infected neighbors has exactly a infected neighbors is given by the binomial distribution as $\binom{d}{a}\alpha_n^a(1-\alpha_n)^{d-a}$. Implicit to this binomial expression is the assumption that nodes associate randomly with other nodes in a network, *or the homogeneous mixing assumption.* This is a common assumption in modeling SIS dynamics and is discussed in [59]. We can now characterize the transition probabilities of the entire population process $\bar{x}_n(d)$, whose sample path evolves according to (2.5). The transition probability from susceptible to infected is:

$$\begin{aligned}\mathbb{P}\left(\bar{x}_{n+1}(d) = \bar{x}_n(d) + \frac{1}{M(d)}\right) &= \sum_{a=0}^{d} \bar{P}_{21}(d,a)\mathbb{P}\left(a \text{ out of } d \text{ neighbours infected}\right) \\ &= \sum_{a=0}^{d} \bar{P}_{21}(d,a)\binom{d}{a}\alpha_n^a(1-\alpha_n)^{d-a}.\end{aligned} \tag{2.9}$$

The transition probability from infected to susceptible is evaluated similarly as:

$$\begin{aligned}\mathbb{P}\left(\bar{x}_{n+1}(d) = \bar{x}_n(d) - \frac{1}{M(d)}\right) &= \sum_{a=0}^{d} \bar{P}_{12}(d,a)\mathbb{P}\left(a \text{ out of } d \text{ neighbours infected}\right) \\ &= \sum_{a=0}^{d} \bar{P}_{12}(d,a)\binom{d}{a}\alpha_n^a(1-\alpha_n)^{d-a}.\end{aligned} \tag{2.10}$$

With the transition probabilities (2.9) and (2.10), define the fractional change of infected and susceptible nodes in the population as:

$$P_{21}(d,\bar{x}_n(d)) \triangleq (1-\bar{x}_n(d))\mathbb{P}\left(\bar{x}_{n+1}(d)=\bar{x}_n(d)+\frac{1}{M}\right) \tag{2.11}$$

$$P_{12}(d,\bar{x}_n(d)) \triangleq \bar{x}_n(d)\mathbb{P}\left(\bar{x}_{n+1}(d)=\bar{x}_n(d)-\frac{1}{M}\right) \tag{2.12}$$

where $\bar{P}_{12}, \bar{P}_{21}$ are defined in (2.4). Here (2.11) is interpreted as the fraction of susceptible nodes $(1-\bar{x}_n(d))$ that become infected $(\bar{P}_{21})$ and (2.12), the fraction of infected nodes $(\bar{x}_n(d))$ that become susceptible $(\bar{P}_{12})$.

Below we formalize the mean field approximation and show that the deviation between the approximation and the true state satisfies an exponential bound, given a time horizon that is of the same order as the size of the network.

Define the following 2− dimensional simplices for $d = 1, 2, \ldots, D$,

$$\mathcal{S}^d = \left\{ \begin{bmatrix} \bar{x}_n(d) \\ 1-\bar{x}_n(d) \end{bmatrix} : \bar{x}_n(d) \in [0,1] \right\}$$

and the following product space, $\mathcal{S} = \mathcal{S}^1 \times \mathcal{S}^2 \times \ldots \times \mathcal{S}^D$.

With $\mathbf{1}$ denoting the D dimensional vector of ones, let $\widetilde{\mathbf{x}}_n = [\bar{x}_n', \ (\mathbf{1}-\bar{x}_n)']$,

$$\begin{aligned}\mathcal{P}_{21}(\widetilde{\mathbf{x}}_n) = [&P_{21}(1,\bar{x}_n(1)), \ \ldots, P_{21}(D,\bar{x}_n(D)), \ P_{12}(1,\bar{x}_n(1)),\\ &\ldots, \ P_{12}(1,\bar{x}_n(2)),\ldots, \ P_{12}(D,\bar{x}_n(D))]\\ \mathcal{P}_{12}(\widetilde{\mathbf{x}}_n) = [&P_{12}(1,\bar{x}_n(1)), \ \ldots, P_{12}(D,\bar{x}_n(D)), \ P_{21}(1,\bar{x}_n(1)),\\ &\ldots, \ P_{21}(1,\bar{x}_n(2)),\ldots, \ P_{21}(D,\bar{x}_n(D))].\end{aligned} \tag{2.13}$$

Theorem 1 (Mean Field Dynamics). *1. The population dynamics Markov chain $\widetilde{\boldsymbol{x}}_n$ evolves according to the following martingale difference driven stochastic difference equation:*

$$\widetilde{\boldsymbol{x}}_{n+1} = \widetilde{\boldsymbol{x}}_n + \frac{1}{M}[\mathcal{P}_{21}(\widetilde{\boldsymbol{x}}_n) - \mathcal{P}_{12}(\widetilde{\boldsymbol{x}}_n)] + \zeta_n. \tag{2.14}$$

Here ζ_n is a $2D$-dimensional martingale difference process[8] *with $\|\zeta_n\|_2 \leq \frac{\Gamma}{M}$ for some positive constant Γ.*

2. *Consider the following deterministic mean field dynamics process associated with the population state:*

$$\boldsymbol{x}_{n+1} = \boldsymbol{x}_n + \frac{1}{M}[\mathcal{P}_{21}(\boldsymbol{x}_n) - \mathcal{P}_{12}(\boldsymbol{x}_n)] \tag{2.15}$$

where $\boldsymbol{x}_n \in \mathcal{S},\ \boldsymbol{x}_n = [x_n',\ (1-x_n)'],\ \mathcal{P}_{21}(\boldsymbol{x}_n)$ and $\mathcal{P}_{12}(\boldsymbol{x}_n)$ are defined as in (2.13) and $\boldsymbol{x}_0 = \tilde{\boldsymbol{x}}_0$.

Then for a time horizon of T points, the deviation between the mean field dynamics $\boldsymbol{x}_n$ in (2.15) and actual population state $\tilde{\boldsymbol{x}}_n$ in (2.14) satisfies

$$\mathbb{P}\Big\{\max_{0\leq n\leq T} \Big\|\boldsymbol{x}_{n+1} - \tilde{\boldsymbol{x}}_{n+1}\Big\|_\infty \geq \epsilon\Big\} \leq C_1 \exp(-C_2\epsilon^2 M)$$

for some positive constants C_1 and C_2 providing $T = O(M)$[9]*.*

Theorem 1 says that for a time horizon of T points, the deviation between the mean field dynamics (2.15) and actual population state (2.14) satisfies an exponential bound. Since the exponential bound $\sum_M \exp(-C\epsilon^2 M)$ is summable, the Borel-Cantelli lemma applies. This in turn implies that, if the deterministic population flow remains forever in some subset of the state space, then the stochastic process will remain in the same subset space with a probability arbitrarily close to one, provided that the population is large enough, see [12].

The proof of Theorem 1 given in the appendix, is inspired by the proof in [12] for approximating stochastic evolution in games. An important feature of the mean field dynamics is that it has a state of dimension D compared to the intractable state dimension $\prod_{d=1}^{D}(M(d)+1)$ of the infected population state $\bar{x}_n$[10], with $\bar{x}_n(d) \in \{0, \frac{1}{M}, \frac{2}{M}, \ldots, \rho(d)\}$.

For the purposes of this thesis, the key outcome of Theorem 1 is that the mean field

[8] ζ_n is a martingale difference process if $\mathbb{E}\{\zeta_n|\mathcal{F}_{n-1}\} = 0$, where $\mathcal{F}_{n-1}$ denotes the sigma-algebra generated by $\{\tilde{\mathbf{x}}_0,\ \tilde{\mathbf{x}}_1,\ \cdots, \tilde{\mathbf{x}}_{n-1}\}$.

[9] Here $\|\mathbf{x}\|_\infty = \max_i |\mathbf{x}|$ denotes the l_∞ norm of vector $\mathbf{x}$.

[10] Note that here $\bar{x}_n$ is scaled by $\rho(d)$.

system $\mathbf{x}_n$ has polynomial dynamics, where for every $d \in \{1, 2, \ldots, D\}$ we have for the infected population state,

$$x_{n+1}(d) = x_n(d) + \frac{1}{M}\left[P_{21}(d, x_n(d)) - P_{12}(d, x_n(d))\right]. \qquad (2.16)$$

These polynomial dynamics will be exploited in chapter 3 for tracking the infected population state.

2.3 Sampling

To track an infection we must have some method by which to learn about the underlying state of the system. The SIS model derived earlier in this chapter informs us on the approximate trajectory, however, for finite networks, there is state noise which causes the infected population state to drift. The state noise in this example arises because even if we know the exact network size, structure, and have estimates of the transmission probabilities that correspond exactly to the true transmission probabilities, the interactions between users are probabilistic. We may know how frequently an individual interacts with their neighbours, but may not know with whom they interact on a particular Sunday, or if they sneeze during this interaction. We encode this shortcoming in of knowledge in the probabilistic model. Even knowing the true probabilities, we do not know a priori, except in trivial cases with transmission probabilities of 0 and 1, if an infection will spread from one user to another at a particular time. We therefore seek to observe the network and learn about the infected population state through sampling.

We consider sampling the social network (2.1). For social networks with large numbers of nodes, it is prohibitive to query each node. This necessitates choosing a sampling methodology in order to estimate the infected population state x from a subset of the total nodes. Each sampled node is asked if it is infected or not and the reply (measurement) noted. Below, we consider two popular methods for sampling large networks, see [3, 16, 20, 30, 38, 39, 49, 54] for an overview of these methods:

2.3.1 Uniform Sampling

When a list of all members of the network is available, but perhaps have no precise information of network structure, we can use uniform sampling to estimate the infected population state. We can obtain an estimate of the infected population state from sampling a sequence of ν individuals uniformly from the population. For each degree, $\nu(d)$ are sampled generating a sequence of independent observations denoted by $\{m_l, l \in \{1, 2, \dots, \nu(d)\}\}$ from the population $M(d)$. The estimate at time n is then given by

$$\hat{x}_n(d) = \frac{1}{\nu(d)} \sum_{l=1}^{\nu(d)} \mathcal{I}(s_n^{(m_l)} = 1). \tag{2.17}$$

At each time n, $\hat{x}_n$ can be viewed as noisy observation of the infected population state x_n.

2.3.2 MCMC Based Respondent-Driven Sampling (RDS)

Respondent-driven sampling (RDS) was introduced by Heckathorn [38, 39] as an approach for sampling from hidden populations in social networks and has gained enormous popularity in recent years. In RDS sampling, current sample members recruit future sample members. RDS sampling does not require knowledge of the structure of the target population. This is important as it allows observation of hidden populations. These populations may be hidden by choice, such as sex workers and the MSM community[11].

The RDS procedure is as follows: A small number of people in the target population serve as seeds. After participating in the study, the seeds recruit other people they know through the social network in the target population. The sampling continues according to

[11] MSM is the health community's term for men who have sex with men and is a classic example of a hidden social network. Members of the social network may know other members, but they may be reluctant to identify themselves to researchers due to the danger or prejudice involved in being known as a member of the community.

this procedure with current sample members recruiting the next wave of sample members until the desired sampling size is reached. In real world applications, this recruitment is typically driven by monetary incentives for participants who successfully recruit their peers.

RDS can be viewed as a form of Markov Chain Monte Carlo (MCMC) sampling (see [33] for an excellent exposition). Let $\{m_l, l \in \{1, 2, \ldots, \nu(d)\}\}$ be the realization of an aperiodic irreducible Markov chain with state space $M(d)$ comprising of nodes of degree d. This Markov chain models the individuals of degree d that are sampled, namely, the first individual m_1 is sampled and then recruits the second individual m_2 to be sampled, who then recruits m_3 and so on. Instead of the independent sample estimator (2.17), an asymptotically unbiased MCMC estimate is then generated as

$$\frac{\sum_{l=1}^{\nu(d)} \frac{\mathcal{I}(s_n^{(m_l)}=1)}{\pi(m_l)}}{\sum_{l=1}^{\nu(d)} \frac{1}{\pi(m_l)}} \tag{2.18}$$

where $\pi(m)$, $m \in M(d)$, denotes the stationary distribution of the Markov chain. For example, a reversible Markov chain with prescribed stationary distribution is straightforwardly generated by the Metropolis Hastings algorithm.

In RDS, the transition matrix and, hence, the stationary distribution π in the estimate (2.18) is specified as follows: Assume that edges between any two nodes i and j have symmetric weights W_{ij} (i.e., $W_{ij} = W_{ji}$, equivalently, the network is undirected). Node i recruits node j with transition probability $W_{ij}/\sum_j W_{ij}$. Then, it can be easily seen that the stationary distribution is $\pi(i) = \sum_{j\in V} W_{ij}/\sum_{i\in V, j\in V} W_{ij}$. Using this stationary distribution, along with the above transition probabilities for sampling agents in (2.18), yields the RDS algorithm.

Since a Markov chain over a non-bipartite connected undirected network is aperiodic [85], the initial seed for the RDS algorithm can be picked arbitratily, and the above estimator is asymptotically unbiased [33].

The key outcome of Sec. 2.3 is that by the central limit theorem (for an irreducible aperiodic finite state Markov chain), the estimate of the probability that a node is infected in a large population (given its degree) is Gaussian. For a sufficiently large number of samples, observation of the infected population state is approximately Gaussian, and the sample observations can be expressed as

$$y_n = Cx_n + v_n \tag{2.19}$$

where $v_n \sim \mathscr{N}(\mathbf{0}, \mathbf{R})$ is the observation noise with the covariance matrix $\mathbf{R}$ and observation matrix C dependent on the sampling process and $x_n \in \mathbb{R}^D$ is the infected population state which evolves according to the polynomial dynamics (2.16).

2.4 Summary

This chapter presented a formulation of a social network and infection diffusion dynamics over that social network. Mean field theory was then utilized to approximate the discrete time diffusion dynamics with an ordinary differential equation. This SIS model and mean field approximation incorporate information about the degree distribution of the underlying network, and so the infection dynamics depend on the underlying network structure. Furthermore, the mean field dynamics are polynomial in the infected population state, a property which will be utilized in the Bayesian filtering algorithm of Chapter 3. Finally, uniform and respondent-driven sampling were presented as methods of generating observations of the infected population state.

Chapter 3

Tracking Infection and Filtering Bounds

In Sec. 2.2, we formulated the mean field dynamics for the infected population state as a system with polynomial dynamics (2.16) and linear Gaussian observations (2.19) obtained by sampling the network. In this section, we consider Bayesian filtering of the infected population state for large networks.

The practical applications of knowing the infected probability state include better estimating the trajectories of social and infectious epidemics. Consider an infection on a network of known degree distribution. Again, we consider a viral infection in a real world social network, such as the Ebola epidemic that occurred in West Africa from 2013-2016. It may be expensive or prohibitively dangerous to generate large and frequent samples of the population to learn about the infection and how it is spreading, but the high level structure of the social network itself, such as the network degree distribution, may be approximately known (consider that the social network in the infected areas would likely have a similar underlying network structure to uninfected areas). We could use expert knowledge to estimate transmission matrices P_{12} and P_{21}. Using these pieces of information we could perform smaller samples over time and use the filtering techniques outlined in this section to generate an estimate that is as accurate as much larger samples. Thus the ability to filter the sampled information and determine the infected probability state saves time, money and decreases the amount of exposure surveyors will have with infected individuals.

It may similarly by difficult to sample on online social networks to learn about technological adoption or information diffusion. Online social networks are massive, and response rates to petitions by interested parties are often very low. What is more, individuals become quickly apathetic to barrages of requests online. Responses are therefore more likely if they can be incentivised, but incentives dramatically increase the cost per sample request. Thus we are again motivated to sample less frequently, and with smaller samples. It is through filtering that we can do this, while maintaining or surpassing the accuracy of observations of population states generated by larger samples.

3.1 Optimal Filtering of Infected Populations

We first describe how to express the mean field dynamics (2.16) in a form amenable to employing the non-linear filter described in [40]. The non-linear filter in [40] can be used with state spaces with polynomial state space evolutions and linear observations, both of which may have Gaussian noise. One of the benefits of integrating domain knowledge of the state space evolution of our system into a dynamical mean field equation is that this equation has polynomial dynamics. That is that the mean field equation 2.16 can be expressed, $x_{n+1} = f(x_n)$, where $f(\cdot)$ is a polynomial function. Once the state space has been thusly expresed, we can directly implement the filter, here described in Sec. 3.1.2.

3.1.1 Mean Field Polynomial Dynamics

Consider a D-dimensional polynomial vector $f(x) \in \mathbb{R}^D$:

$$f(x) = A_0 + A_1 x + A_2 xx' + A_3 xxx' + \dots \tag{3.1}$$

where the co-efficients $A_0, A_1, \ldots, A_i$ denote rank $1, 2, \ldots, (i+1)$ tensors. $A_i xx \ldots x'$ is a vector with r^{th} entry given by

$$A_i xx \ldots x'(r) = \sum_{j_1, j_2, j_3, \ldots, j_i} A_i(r, j_1, j_2, \ldots, j_i) x_{j_1} x_{j_2} \ldots x_{j_i}$$

where $A_i(r, j_1, j_2, \ldots, j_i)$ is the $r, j_1, j_2, \ldots, j_i$ entry of tensor A_i and x_j is the j^{th} entry of x.

Equation (2.16) has polynomial dynamics and can be expressed in the form of (3.1) by constructing the tensors A_i from $\bar{P}_{12}$, $\bar{P}_{21}$ and $\rho(d)$. We illustrate how to express the mean field dynamics of (2.16) in the form (3.1) for[12] $d = 2$. First, we note the average degree is $\sum_{d=1}^{D} d\,\rho(d)$ and the link probability given in (2.8) can be expressed as $\alpha_n = \phi' x_n$, where ϕ is defined as:

$$\phi = \left[\frac{\rho(1)}{\sum_{d=1}^{D} d\,\rho(d)}, \frac{2\rho(2)}{\sum_{d=1}^{D} d\,\rho(d)}, \ldots, \frac{D\rho(D)}{\sum_{d=1}^{D} d\,\rho(d)} \right]'.$$

Mean field dynamics for $d = 2$ in (2.16) is given as:

$$x_{n+1}(2) = x_n(2) + \frac{1}{M} \left[P_{21}(2, x_n) - P_{12}(2, x_n) \right]. \tag{3.2}$$

For all terms with factor $\bar{P}_{12}$ there is a corresponding term with factor $\bar{P}_{21}$, so for convenience we will account for all of the former with Ω and the latter with Ω^*.

$$x_{n+1}(2) = x_n(2) + \Omega - \Omega^* \tag{3.3}$$

[12] The procedure is same for all degrees.

where Ω is all the terms in an expanded (2.16), with $d = 2$, containing $\bar{P}_{12}$ and

$$\begin{aligned}\Omega = \theta \Bigg[&\frac{1}{M}\bar{P}_{12}(2,0)(\phi' x_n)^2 + \frac{2}{M}\bar{P}_{12}(2,1)(\phi' x_n) - \frac{2}{M}\bar{P}_{12}(2,1)(\phi' x_n)^2 + \frac{1}{M}\bar{P}_{12}(2,2) \\ &- \frac{2}{M}\bar{P}_{12}(2,2)(\phi' x_n) + \frac{1}{M}\bar{P}_{12}(2,2)(\phi' x_n)^2 - \frac{x_n}{M}\bar{P}_{12}(2,0)(\phi' x_n)^2 - \frac{2x_n}{M}\bar{P}_{12}(2,1)(\phi' x_n) \\ &+ \frac{2x_n}{M}\bar{P}_{12}(2,0)(\phi' x_n)^2 - \frac{x_n}{M}\bar{P}_{12}(2,2) + \frac{2x_n}{M}\bar{P}_{12}(2,2)(\phi' x_n) - \frac{x_n}{M}\bar{P}_{12}(2,2)(\phi' x_n)^2 \Bigg]. \end{aligned} \quad (3.4)$$

By grouping all of the terms of (3.3) by their order in $x_n(d)$, we can generate the tensors of (3.1). The contributions to the tensors of (3.1) by Ω are:

$$\begin{aligned} A_0(2) &= \frac{\theta \bar{P}_{12}(2,2)}{M} \\ A_1(2,:) &= \phi \left[\frac{2\theta(\bar{P}_{12}(2,1) - \bar{P}_{12}(2,0))}{M} \right] \\ A_2(2,:,:) &= \phi\phi' \left[\frac{\theta(\bar{P}_{12}(2,0) - 2\bar{P}_{12}(2,1) + \bar{P}_{12}(2,2))}{M} \right] \\ A_2(2,2,:) &= \phi \left[\frac{2\theta(\bar{P}_{12}(2,1) - \bar{P}_{12}(2,0))}{M} \right] \\ A_3(2,2,:,:) &= -\phi\phi' \left[\frac{\theta(\bar{P}_{12}(2,0) - 2\bar{P}_{12}(2,1) + \bar{P}_{12}(2,2))}{M} \right] \end{aligned} \quad (3.5)$$

where $A_i(j_1, j_2, \ldots, j_{i-1}, :, :)$ is a submatrix of tensor A_i and $A_i(j_1, j_2, \ldots, j_i, :)$ is a subvector of tensor A_i. By following (3.5) for Ω and Ω^* for all d, we can generate all the coefficients in the tensors of (3.1) from $\bar{P}_{12}$, $\bar{P}_{21}$, and $\rho(d)$. We note that the polynomial that defines the dynamics of the network is of order $D^* + 1$, where D^* is the highest degree node with complex dynamics, i.e: $\bar{P}_{21}(d,a) = \bar{P}_{12}(d,a) = \kappa$ for all $d > D^*$ and for all a, where κ is constant with respect to both d and a.

3.1.2 Optimal Filter for Polynomial Dynamics

Optimal Bayesian filtering refers to recursively computing the conditional mean estimate $\mathbb{E}\{x_k|Y_k\}$, for $k = 1, 2, \cdots$ where Y_k denotes the observation sequence $y_1, \ldots, y_k$. (The term optimal refers to the fact that the conditional mean estimate is the minimum variance estimate). To estimate the infected population state using the sampled observations (2.19), we employ an optimal Bayesian filter recently developed for polynomial systems. In general for nonlinear or non-Gaussian systems, there is no finite dimensional filtering algorithm; however, it is shown in [40] that for Gaussian systems with polynomial dynamics, one can devise a finite dimensional filter (based on the Kalman filter) to compute the conditional mean estimate.

Below we present only the relevant terms in the model that are used in the filtering equations, see [40] for details. The non-linear filter prediction and update equations are given as:

Prediction step:

$$\begin{aligned}
\hat{x}_n^- &= \mathbb{E}[x_n|Y_{n-1}] = \mathbb{E}[f(x_{n-1})|Y_{n-1}] \\
H_n^- &= \mathbb{E}[(x_n - \hat{x}_n)(x_n - \hat{x}_n)'|Y_{n-1}] \\
&= \mathbb{E}[(f(x_{n-1}) - \mathbb{E}[f(x_{n-1})|Y_{n-1}] + v_{n-1}) \times (f(x_{n-1}) - \mathbb{E}[f(x_{n-1})|Y_{n-1}] + v_{n-1})'|Y_{n-1}] \\
&= \mathbb{E}[f(x_{n-1})f(x_{n-1})'|Y_{n-1}] - \mathbb{E}[f(x_{n-1})|Y_{n-1}] \times \mathbb{E}[f(x_{n-1})|Y_{n-1}]' + \mathbf{Q}_{n-1}
\end{aligned} \tag{3.6}$$

where $\hat{x}_n^-$ denotes the priori expectation of the state vector (infected population state) x_n; $Y_n = \{Y_{n-1}, y_n\}$ denotes the observation process; H_n^- denotes the priori state co-variance estimate at time n; and v_n denotes the Gaussian random variable representing the state noise at time n, with covariance $\mathbf{Q}_n$.

H_n^- and $\hat{x}_n$ are initialized with some initial estimate $\hat{x}_0$ and estimate covariance H_0^-. The filter relies upon being able to compute the expectation $\mathbb{E}[f(x_{n-1})f'(x_{n-1})|Y_{n-1}]$ in terms of $\hat{x}_{n-1}$ and H_n^-. When $f(\cdot)$ is a polynomial, $f(x_{n-1})f(x_{n-1})'$ is a function of x_{n-1}, and the conditional expectations in (3.7) can be expressed only in terms of $\hat{x}_{n-1}$ and H_n^-, permitting a closed form[13] prediction step.

Update step:

$$\begin{aligned}
\hat{x}_n^+ &= \mathbb{E}[x_n|Y_n] = \hat{x}_n^- + H_n^- C'(\mathbf{R}_n + CH_n^- C')^{-1}(y_n - C\hat{x}_n^-) \\
K_n &= H_n^- C'(\mathbf{R}_n + CH_n^- C')^{-1} \\
H_n^+ &= (I - K_n C)H_n^-(I - K_n C)' + K_n \mathbf{R}_n K_n'
\end{aligned} \tag{3.7}$$

where $\hat{x}_n^+$ denotes the posteriori expectation of the state vector; H_n^+ denotes the posteriori state co-variance estimate at time n; C denotes the state observation matrix; $\mathbf{R}_n$ denotes the observation noise co-variance matrix; K_n denotes the filter gain; and I denotes the identity matrix.

Since the dynamics of (2.16) are polynomial, the prediction and update steps of (3.6) and (3.7) can be implemented without approximation. These expressions constitute the optimal non-linear filter and can be used to track the evolving infected population state.

3.2 Posterior Cramèr-Rao Bounds

The PCRLB provides a theoretical performance limit for an unbiased estimator in a nonlinear filtering problem [79]. While not of direct practical application, the PCRLB provides an understanding of the fundamental limits of our filtering algorithm, and provides a method by which we can investigate the properties of a network. The PCRLB can, for instance, inform us on the difference in performance limit of different network types, or under different sampling schemes. Understanding the PCRLB therefore allows us to better design our data gathering methodologies. In PCRLB, the Fisher Information Matrix (FIM) is derived by evaluating the expectation with respect to the joint distribution of the measurements and the system states up to the current time. It is determined only by model parameters and

[13] For an explicit implementation of such a filter for a third order system with an exact priori update equation for H_n^- and $\hat{x}_n^-$, see [40].

the initial state and is thus independent of any specific realization of the system state.

Below we compute the PCRLB for the polynomial dynamical system (2.16) with linear observations with Gaussian noise (2.19). We first consider the dynamical system with mean zero Gaussian state noise having the covariance matrix $\mathbf{Q}$, then formulate the PCRLB for the polynomial dynamical system (2.16) without any state noise.
Let $X_n = \{X_{n-1}, x_n\}$ denote the state sequence at time n and J_n denote the $((n+1)D \times (n+1)D)$ Fisher information matrix[14] of X_n [79]. The following recursion is used to evaluate J_n in [79]. Let $\mathbb{P}(X, Y) = p_n$ denote the joint distribution and $\Delta_x^y = \nabla_x \nabla_y'$ denote the vector differential operator. Then [79],

$$J_n \;=\; \mathbb{E}\{-\Delta_{x_n}^{x_n} \log(p_n)\} \;-\; \mathbb{E}\{-\Delta_{x_n}^{X_{n-1}} \log(p_n)\}[\mathbb{E}\{-\Delta_{X_{n-1}}^{X_{n-1}} \log(p_n)\}]^{-1}\mathbb{E}\{-\Delta_{X_{n-1}}^{x_n} \log(p_n)\}$$

and corresponds to the inverse of the $D \times D$ right lower block of $[J_n]^{-1}$. The recursion for J_n is given by:

$$J_{n+1} = \Lambda_n^{22} - \Lambda_n^{21}\left(J_n + \Lambda_n^{11}\right)^{-1}\Lambda_n^{12}$$

where

$$\begin{aligned}
\Lambda_n^{11} &= \mathbb{E}\{(\nabla_{x_n} f_n'(x_n))\mathbf{Q}_n^{-1}(\nabla_{x_n} f_n'(x_n))'\} \\
\Lambda_n^{12} &= \mathbb{E}\{\nabla_{x_n} f_n'(x_n)\}\mathbf{Q}_n^{-1} \\
\Lambda_n^{21} &= \{D_n^{12}\}' \\
\Lambda_n^{22} &= \mathbf{Q}_n^{-1} + C\mathbf{R}_n^{-1}C'
\end{aligned} \tag{3.8}$$

where C is the linear observation matrix and $\mathbf{R}$ is the observation noise covariance matrix

[14] The error covariance $\mathscr{P}$ is,

$$\mathscr{P} = \mathbb{E}\{(\hat{X} - X)(\hat{X} - X)'\} \geq J^{-1}$$

where J is the Fisher Information matrix, X is the state, $\hat{X}$ is the state estimate, and Y measured data. The elements of the Fisher information matrix J are given by:

$$J_{ij} = \mathbb{E}\left(\frac{\partial^2 \log p_{x,y}(X, Y)}{\partial X_i \partial X_j}\right)$$

of (2.19) and $\nabla_{x_n} f'_n(x_n)$ is given as:

$$\nabla_{x_n} f'_n(x_n) = \left[\frac{\partial f}{\partial x_n(0)}, \frac{\partial f}{\partial x_n(1)}, \ldots, \frac{\partial f}{\partial x_n(D)}\right]'$$

$$\frac{\partial f}{\partial x_n(0)} = 0 + \frac{\partial}{\partial x_n(0)}\left[A_1 x_n\right] + \frac{\partial}{\partial x_n(0)}\left[A_2 x_n x'_n\right] + \frac{\partial}{\partial x_n(0)}\left[A_3 x_n x_n x'_n\right] \tag{3.9}$$

Thus

$$\nabla_{x_n} f'_n(x_n) = A_1 + (A_2 + A'_2)x_n + (A_{3_{ijk}} + A_{3_{jki}} + A_{3_{kji}})x_n x'_n \tag{3.10}$$

where $A_{3_{ijk}}$ indicates the ordering of indices of the tensor A_3 and is analogous to a higher dimensional transpose.

When there is no state noise in the system evolution, as in the mean field dynamics equation (2.16), to compute the PCRLB, we perturb the state evolution in (2.16) with pairwise independent Gaussian random vectors having covariance matrix $\mathbf{Q}_\epsilon = \epsilon I$, replacing the singular state evolution by a perturbed system $p_\epsilon(x_{n+1}|x_n)$. For the purturbed system we have,

$$-\log p_\epsilon(x_{n+1}|x_n) = c + \frac{1}{2}\{x_{n+1} - f_n(x_n)\}'\mathbf{Q}_\epsilon^{-1}\{\bar{x}_{n+1} - f_n(x_n)\} \tag{3.11}$$

where c is a constant. The recursion for J_n is then given by:

$$J_{n+1} = \Lambda^{22}_{\epsilon,n} - \Lambda^{21}_{\epsilon,n}\left(J_n + \Lambda^{11}_{\epsilon,n}\right)^{-1}\Lambda^{12}_{\epsilon,n}$$

where

$$\begin{aligned}
\Lambda^{11}_{\epsilon,n} &= \frac{1}{\epsilon}\mathbb{E}\{(\nabla_{x_n} f'_n(x_n))(\nabla_{x_n} f'_n(x_n))'\} \\
\Lambda^{12}_{\epsilon,n} &= \frac{1}{\epsilon}\mathbb{E}\{\nabla_{x_n} f'_n(x_n)\} \\
\Lambda^{21}_{\epsilon,n} &= \{\Lambda^{12}_n\}' \\
\Lambda^{22}_{\epsilon,n} &= \frac{1}{\epsilon}I + C\mathbf{R}_n^{-1}C'
\end{aligned} \tag{3.12}$$

PCRLB: Erdős-Rényi vs Scale-Free

We used the above formulas to compute the PCRLB for the mean field dynamics model (2.16), under two simulation schemes.

Scheme A.) Degree distribution based: The deterministic mean field evolution given in (2.16) is simulated with linear Gaussian observations (2.19) and zero state noise. The infection transition probabilities $\bar{P}_{12}$ and $\bar{P}_{21}$ were chosen uniformly at random over $[0, 1]$. The observation noise covariance matrix $\mathbf{R}$ is a random positive definite matrix[15] with entries in $[0, 10^{-6}]$. We note that the diffusion parameter $\theta = 1$, observation matrix $C = I$, and maximum degree $D = 20$ unless otherwise stated.

Scheme B.) Network based: The network based simulation consisted of generating a network, propagating an infection according to Sec.2.1.2 over this network, and then finally sampling that network using the uniform sampling mechanism described in Sec.2.3. For the network simulations, scale-free networks of $M = 10000$ nodes were generated using the Barabasi-Albert model and Erdős-Rényi networks of $M = 10000$ were generated by randomly creating links between nodes. The infection was initialized by infecting nodes at time $n = 0$ with probability 0.01. This infection was propagated for 10000 timesteps. At each timestep 5000 samples were obtained according to the uniform sampling described in Sec.2.3 to generate an observation at that timestep. These observations were then filtered. State and observation covariance matrices $\mathbf{Q}$ and $\mathbf{R}$ used in filtering were computed empirically from the data. State and observation noise covariance matrices were estimated to be constant for all time steps n, that is $\mathbf{Q}_n = \mathbf{Q}$ and $\mathbf{R}_n = \mathbf{R}$ for all n.

We computed the PCRLB for two different network types:

Type a.) Scale-free network with degree distribution $\rho(d) \propto d^{-\gamma}$ where $\gamma = 2.7$. Such networks

[15] One way is to generate a sample from the probability measure on real symmetric positive-definite matrices such as Inverse-Wishart.

arise in online social networks such as Twitter [51] and in the link network of the World Wide Web [2].

Type b. Erdős-Rényi network with degree distribution $\rho(d) \propto \frac{e^{-\lambda d}}{d!}$ where $\lambda = 2.7$. [16]

Fig. 3.1 displays the PCRLB and mean square error of the degree based simulations. Interestingly, it can be seen from Fig. 3.1 that both PCRLB and its slope are insensitive to the underlying network structure when observation noise covariance, $\mathbf{R}$ in (2.19), is not network dependent. Under the mean field model, differences between PCRLB for different network types can therefore be attributed to either different state or different observation noise covariances. The PCRLB and Bayesian filter estimate mean square error of network based simulations are shown in Fig. 3.2. In Fig. 3.2, a network of 10000 nodes is generated and an infection is propagated through the network over 10000 time steps. This infection is then sampled 5000 times at each timestep to generate observations of the infected population states. State and observation covariances are computed empirically and assumed constant throughout the duration of the system dynamics. The slower convergence of the filter estimate on an Erdős-Rényi when compared to a scale-free network filter estimate corresponds to the Erdős-Rényi network observations having a larger observation noise covariance. It is seen that PCRLB gives a lower bound on the MSE.

3.3 Summary

This Chapter presents a filtering method for tracking the infected population state of a social network. This enables the combination of apriori knowledge of the infection dynamics with observations to better estimate the fraction of infected users in a network. First, we expressed the mean field dynamics of Sec. 2.1.2 in a polynomial form, which is amenable the non-linear Bayesian filtering algorithm described in Sec. 3.1. The non-linear Bayesian algorithm filters observations to produce an optimal estimate of the infected population

[16] The value $\lambda = 2.7$ was chosen since it is similar to the the out-degree of the World Wide Web, see [15]

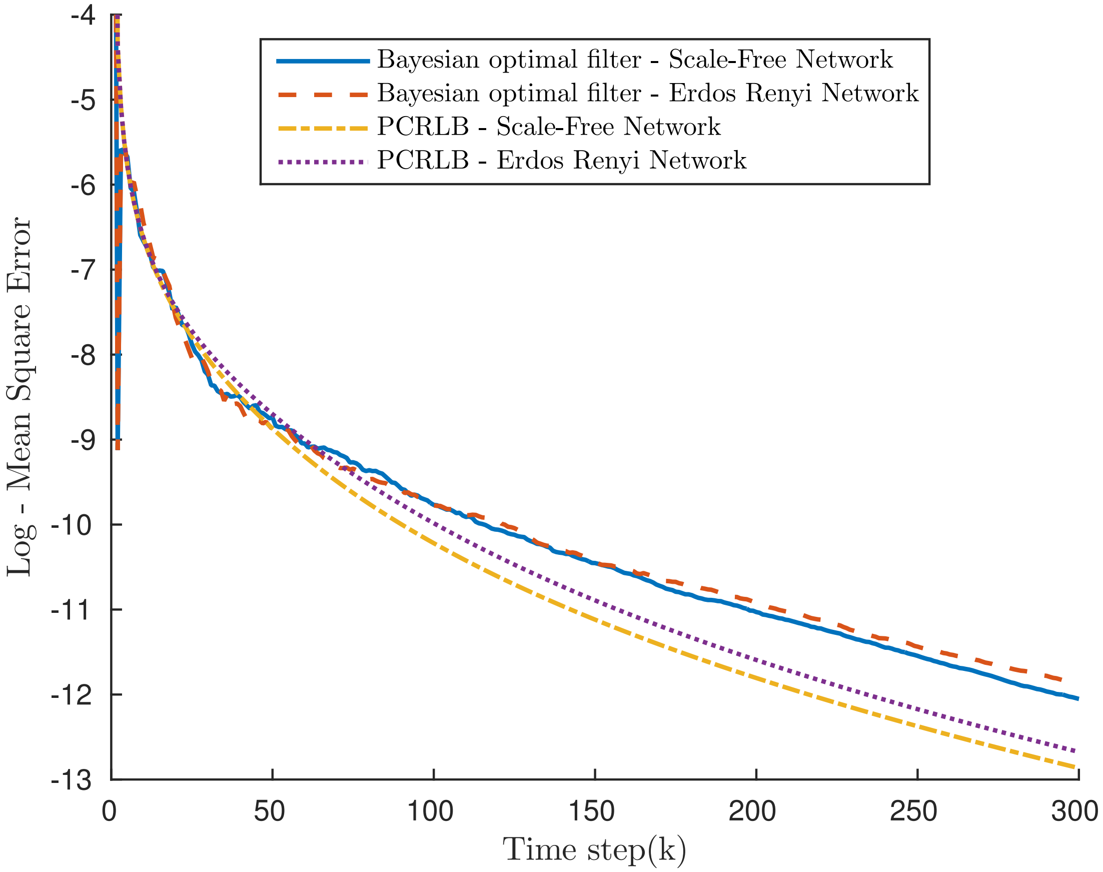


Figure 3.1: Filtered estimate log mean square error and PCRLB for deterministic mean field evolution with scale-free and Erdős-Rényi degree distributions.

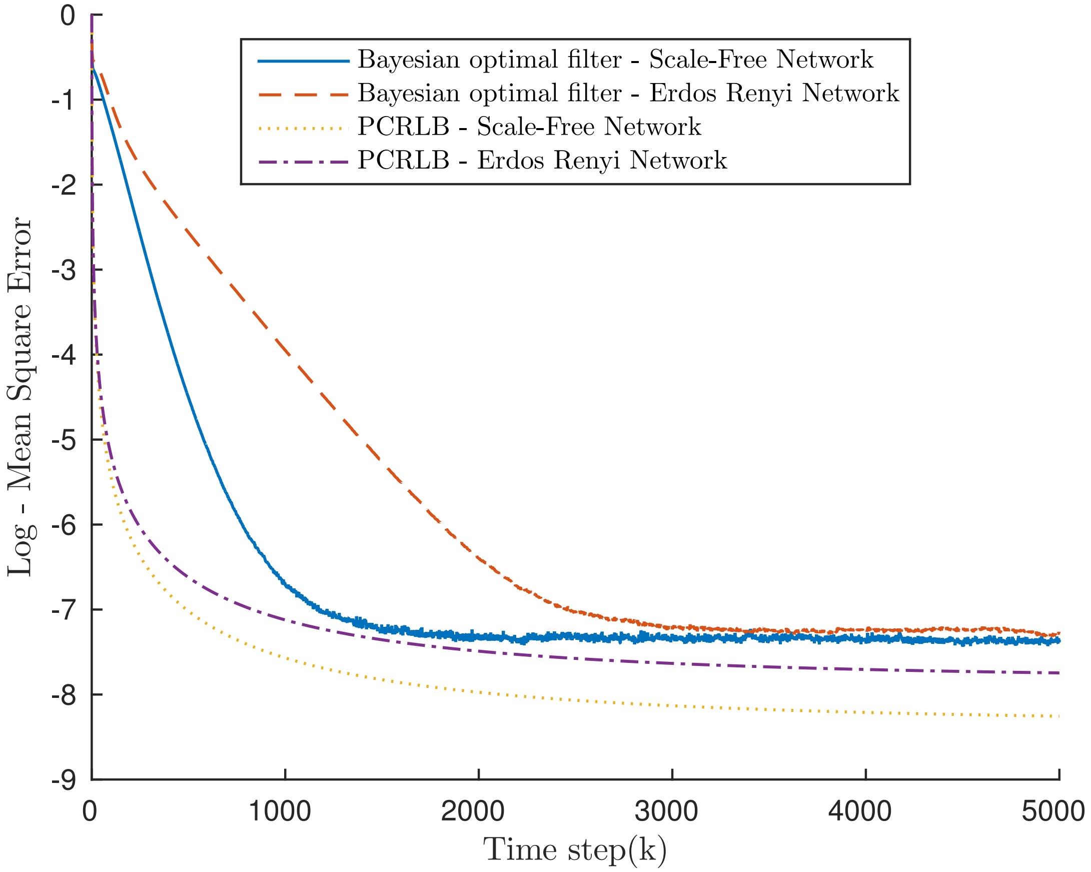


Figure 3.2: Filtered estimate log mean square error and PCRLB of simulated networks - scale-free and Erdős-Rényi.

state. This online estimation tracks the infected population state as it evolves over time. Further, the PCRLB was computed to provide a theoretical performance limit of the filtering problem. The PCRLB was computed for two network types, scale-free networks and Erdős-Rényi networks, and compared.

Chapter 4

Extended Diffusion Models: Homophily and Game Theoretic Dynamics

The SIS model, like any model, is a simplification of a real world phenomena. Infections in the real world operate with far more complexity than can be adequeately summarized by a simple SIS model. To better describe real world infections and opinion diffusion we consider a powerful extension to the SIS model, as well as an alternative model for the underlying infection dynamics. We extend the SIS model to incorporate homophily. Homophily is the tendency for individuals to engage with people similar to themselves [63]. These similarities characterize distinct groups, and these groups may affect how infections spread through a network. For example these distinct groups may be formed by the political affiliations of individuals within the network. This affiliation may then impact how knowledge of a news item diffuses through the network. This incorporation of types and groups into the model does not impede the use of the Bayesian polynomial filter, and thus we can still use this optimal filtering algorithm to track infections in networks with homophily. The alternative model for infection dynamics we present in this thesis is one which inorporates game theoretic interactions into the network dynamics. The Moran process from evolutionary dynamics is adopted as the underlying model for infection (or rather, in this case cooperation) diffusion. The Moran process models the dynamics of a population where individuals belonging to subtypes or substrategies interacts temporally on a stuctured network. This model, when approximated with mean field dynamics, does not generally have a polynomial state evolution, which means the polynomial filter can not be applied without futher approximation. We also consider the pair approximation for the Moran process, in which the state incorporates additional information about the local infection/cooperation structures of the network. The pair approximation does not admit a polynomial state evolution. We therefore utilize a particle filter for tracking the infected/cooperating population state which can be utilized along with the mean field or pair approximations.

4.1 Infected Population Estimation in Homophily Networks

An infection doesn't affect the elderly the same way it affects the young. Minor annoyances for a student athlete can be hospitalizations for the very young or immunocomprimised. Similarly, your engagment with a news article may depend on your prior previous affiliations, and your initial exposure to that article may depend on how many of your friends share your political affiliation or subscribe to different beliefs. This more complex diffusion and network dependence on individual types can be modeled by a network with homophily. In this section we extend the SIS model presented in Sec. 2.2 to include homophily, using the approach in [57].

4.1.1 SIS Model with Homophily

Let the social network be comprised of $\mathcal{B}$ distinct types of individuals[17]. Let μ_b denote the fraction of individuals of type b, for $b \in \{1, 2 \ldots, \mathcal{B}\}$. We define the total number of nodes of type b as

$$M^b = \sum_{m \in V} \mathcal{I}\big(\mathbb{T}(m) = b\big). \tag{4.1}$$

[17] For example these types may be groups characterized by political affiliation, age, or income.

where we define $\mathbb{T}(m)$ to be the type of node m. We define the total number of nodes of degree d and type b as

$$M^b(d) = \sum_{m \in V} \mathcal{I}\big(\Xi^{(m)} = d, \mathbb{T}(m) = b\big). \tag{4.2}$$

The degree distributions of each type can be different from one another, and so we specify the degree distribution of each type b as $\rho^{[b]}(d)$, where

$$\rho^{[b]}(d) = \frac{M^b(d)}{M^b}. \tag{4.3}$$

These different types interact with each other according to the transition matrix Υ, with entries $\varsigma_{bc} = P(\text{individual of type } c \text{ meets individual of type } b)$, and is defined as:

$$\Upsilon = \begin{pmatrix} \varsigma_{11} & \cdots & \varsigma_{1\mathcal{B}} \\ \vdots & \dots & \vdots \\ \varsigma_{\mathcal{B}1} & \cdots & \varsigma_{\mathcal{B}\mathcal{B}} \end{pmatrix} \tag{4.4}$$

We denote by $\bar{x}_{n,b}(d)$ the fraction of agents of type b and degree d that are infected at time n. Recall that in Sec. 2.1.2 we defined the infected link probability, α. Since types of individuals interact with others at different rates, ς, they do not share the same infected link probability. Instead of one infected link probability, we define a vector of infected link probabilities, one for each type:

$$\alpha_b(\bar{x}_n) = \sum_{c=1}^{\mathcal{B}} \varsigma_{bc} \frac{\sum_{d=1}^{D} d\, \rho^{[c]}(d)\, \bar{x}_{n,c}(d)}{\sum_{d=1}^{D} d\, \rho^{[c]}(d)}. \tag{4.5}$$

In an SIS model with homophily, the transition probabilities between infected and susceptible states, first introduced in (2.4), may depend on the type b of the node m. Therefore we define

$$\bar{P}^b_{ij}(d,a) \ = \mathbb{P}\left(s^{(m)}_{n+1} = j | s^{(m)}_n = i, \Xi^{(m)} = d, F^{(m)}_n = a, \mathbb{T}(m) = b\right) \tag{4.6}$$

As in Sec. 2.1.2, with $\alpha_{n,i}$ and $\bar{P}^b_{ij}$, the transition probability from susceptible to infected for an individual m of type b can be evaluated as:

$$\mathbb{P}\left(\bar{x}_{n+1,b}(d) = \ \bar{x}_{n,b}(d) + \frac{1}{M^b(d)}\right) = \sum_{a=0}^{d} \theta \bar{P}_{21}^{b}(d,a) \binom{d}{a} \alpha_{n,b}^{a} (1-\alpha_{n,b})^{d-a}. \tag{4.7}$$

where $M(d,b) = \sum_{m \in V} \mathcal{I}\big(\Xi^{(m)} = d, \mathbb{T}(m) = b\big)$ is the total number of nodes of degree d and type b. Similarly, the transition probability from infected to susceptible is evaluated as:

$$\mathbb{P}\left(\bar{x}_{n+1,b}(d) = \bar{x}_{n,b}(d) - \frac{1}{M^b(d)}\right) = \sum_{a=0}^{d} \theta \bar{P}_{12}^{b}(d,a) \binom{d}{a} \alpha_{n,b}^{a} (1-\alpha_{n,b})^{d-a}. \tag{4.8}$$

With the transition probabilities (4.7) and (4.8), we define the change on the fraction of infected and susceptible nodes in the population as:

$$P_{21}^{b}(d, \bar{x}_{n,b}(d)) \stackrel{\Delta}{=} (1-\bar{x}_{n,b}(d)) \rho^{[b]}(d) \mathbb{P}\left(\bar{x}_{n+1,b}(d) = \bar{x}_{n,b}(d) + \frac{1}{M^b(d)}\right) \tag{4.9}$$

$$P_{12}^{b}(d, \bar{x}_{n,b}(d)) \stackrel{\Delta}{=} \bar{x}_{n,b}(d) \rho^{[b]}(d) \mathbb{P}\left(\bar{x}_{n+1,b}(d) = \bar{x}_{n,b}(d) - \frac{1}{M^b(d)}\right) \tag{4.10}$$

where $\bar{P}_{12}^{b}, \bar{P}_{21}^{b}$ are defined in (4.6). As in Sec. 2.2, we use P_{12}^{b} and P_{21}^{b} along with Theorem 1 to express a mean field system $\mathbf{x_n}$ with polynomial dynamics where for every $d \in \{0,1,2,\ldots,D\}$ and every $b \in \{1,2,\ldots,\mathcal{B}\}$ we have for the infected population state,

$$x_{n+1,b}(d) = x_{n,b}(d) + \frac{1}{M}\left[P_{21}^{b}(d, x_{n,b}(d)) - P_{12}^{b}(d, x_{n,b}(d))\right]. \tag{4.11}$$

The addition of homophily to the SIS model changes the state evolution, but $P_{ij}^{b}(d, x_{n,b}(d))$ of (4.11) and $P_{ij}(d, x_n(d))$ of (2.16) are both polynomial functions of their respective states $x_{n,b}(d)$ and $x_n(d)$. Thus the addition of homophily preserves the polynomial dynamics of (2.16) and we can again exploit this polynomial state evolution to apply the optimal filter of Sec. 3.1.

4.2 Moran process: Game Theoretic Network Dynamics

In this section, we present an alternative model for infection and opinion diffusion on social networks: the Moran process. The Moran process is a model used in mathematical biology to investigate and analyze evolution and the adoption of mutations in a population. In the Moran process, diffusion is dependent on the defined by the expected outcomes of games between individuals on the network. This process is distinguished from the SIS model because the infection/cooperation transmission probabilities may depend on the current infection/cooperation state of the network. We then present mean field approximations and a pair approximation to the Moran process which incorporate network structure through the network's degree distribution, extending upon existing work in evolutionary dynamics which presents mean field and pair approximations without additional network structure information [56] [68].

4.2.1 Moran Process

We consider the Moran process with frequency dependent fitness to model the evolution of cooperation in finite populations [56] [68]. Similar to the discrete SIS model, the Moran process considers a population of constant size M. In the Moran process individuals interact -play a game- with their environment. These individuals can adopt one of two states: they can cooperate (C) with their neighbors or defect (D). Thus the state of each node is binary, as in Sec. 2.2, and at each time instant $n = 0, 1, \ldots$, an individual, or node, m has states

$$s_n^{(m)} \in \{1 \text{ (Cooperating)}, 2 \text{ (Defecting) }\}.$$

The individual is then assumed to receive some payoff which is dependent on their action and the actions of their neighbors. Payoffs for games between two players can be summarized with a 2x2 table, where the row player is the individual chosen for interaction and the column player is a neighbour with which he interacts:

$$\begin{array}{cc} & \begin{array}{cc} C & D \end{array} \\ \begin{array}{c} C \\ D \end{array} & \begin{pmatrix} \mathbf{a} & \mathbf{b} \\ \mathbf{c} & \mathbf{d} \end{pmatrix} \end{array}$$

That is, if an individual chooses to cooperate, and the neighbour with whom he is interacting chooses to also cooperate, he receives a payoff of a. Similarly, if he cooperates and his neighbour defects, he receives a payoff b. The quintessential example of a two-player game is the prisoner's dilemma (PD), a game in which $c > a$ and $d > b$. In the prisoner's dilemma, two criminals are being individually questioned for a crime they committed. The investigators have enough evidence to convict each prisoner for minor sentences, but need testimony for major convictions. To coax each prisoner to betray the the other, the investigators offer a deal: "Squeal on your fellow prisoner and we'll reduce your sentence". Cooperating with your fellow prisoner would then be staying silent, and not squealing on them. We can imagine a payoff matrix which looks like:

$$\begin{array}{cc} & \begin{array}{cc} C & D \end{array} \\ \begin{array}{c} C \\ D \end{array} & \begin{pmatrix} \mathbf{-1} & \mathbf{-5} \\ \mathbf{0} & \mathbf{-3} \end{pmatrix} \end{array}$$

Thus, if both prisoners stay silent each receives a minor sentence, but if one squeals and the other stays silent, then the silent individual is punished heavily and the prisoner who squeals goes free. If both prisoners squeal, both receive a moderate sentence. This scenario is of interest, because the cooperating strategy is dominated by the defecting strategy. Regardless of what the other prisoner does, a rational prisoner should choose to defect to minimize its sentence. This dominant strategy does not, however, minimize jail time for the group. The lowest net jail time is both prisoner's cooperating with each other and staying silent. In the structured Moran process, we model a population which interacts through playing these games.

In the Moran process, the specific neighbour with whom an individual interacts is chosen from amongst its set of neighbours. Under frequency dependent selection, we define the payoff for a cooperator with d neighbors, a cooperating neighbors, and $d - a$ defecting neighbors to be

$$\pi_1^\dagger(d, a) = \mathbf{a} \cdot a + (d - a)\mathbf{b}$$

Similarly, the payoff for a defector with d neighbors, a cooperating neighbors, and $d - a$ defecting neighbors is

$$\pi_2^\dagger(d, a) = \mathbf{c} \cdot a + (d - a)\mathbf{d}.$$

We do not always know the exact neighbourhood of the individuals we are interested in. We thus look to the average fitnesses, given some *infected link probability* or, analogously, *cooperating link probability* α_n:

$$\pi_1(d, \alpha_n) = \mathbf{a} \cdot \alpha_n d + (d - \alpha_n d)\mathbf{b} \qquad \pi_2(d, \alpha_n) = \mathbf{c} \cdot \alpha_n d + (d - \alpha_n d)\mathbf{d}. \tag{4.12}$$

We presume that the fitness of an individual is linear in these payoffs. For some selection strength $w \in [0, 1]$ we have that the fitness of an individual with d neighbours and a cooperating neighbours is:

$$f_1^\dagger(d, \alpha_n) = 1 - w + w\pi_1^\dagger(d, a) \quad , \quad f_2^\dagger = 1 - w + w\pi_2^\dagger(d, a)$$

and a randomly selected cooperating/defecting individual have respective fitnesses:

$$f_1(d, \alpha_n) = 1 - w + w\pi_1(d, \alpha_n) \quad , \quad f_2 = 1 - w + w\pi_2(d, \alpha_n).$$

Note that the standard Moran process is for an unstructured population, that is all individuals interact with all others. This is equivalent to the complete network case, where individuals are arranged in a complete graph structure with $d = N - 1$, where N is the total number of individuals in the network.

Types of Updating

In the Moran process, at each timepoint n an update occurs. The Moran process assumes a constant network size of M individuals, and at each timepoint n an update occurs in which one individual "dies" or becomes open to changing their opinion/adopt a product/become infected and another individual gives "birth" or influences/infects the now open individual. The probability that a particular individual "dies" or "gives birth" may depend on their fitness. We first present two types of updating commonly investigated in evolutionary dynamics [67].

1. death-Birth updating: One individual in the population, m, is selected uniformly at random to die. A new individual is born at the vacant spot and its neighbors compete to influence its strategy $s_{n+1}^{(m)}$. The strategy of the newborn individual is chosen probabilistically from its neighbour's strategies with probability weight according to the fitness of the competing strategies.

$$\bar{P}_{21}(d, a, \alpha_n) = \mathbb{P}\left(s_{n+1}^{(m)} = 1 | s_n^{(m)} = 2, \Xi^{(m)} = d, F_n^{(m)} = a, \alpha_n\right) = \frac{a f_1(d, \alpha_n)}{a f_1(d, \alpha_n) + (d - a) f_2(d, \alpha_n)}. \tag{4.13}$$

2. Birth-death updating: One individual from the population, m', is chosen with probability proportional to their fitness to reproduce. This individual then picks one of its neighbours, m, uniformly at random, and changes the strategy of that neighbour to its own.

   First, we consider the probability that a node m' is picked from the population to reproduce. A randomly chosen node has degree d and a cooperating neighbours with probability $\bar{x}_n(d)\rho(d)\binom{d}{a}\alpha_n^d(1-\alpha_n)^{(d-a)}$ and has fitness $f_2(d,a)$. We can, therefore, express the probability that the reproducing node is a cooperator with degree d and a cooperating neighbours as the fractional fitness of all nodes with degree d and a neighbours over the fitness of the entire network.

$$\mathbb{P}\left(s_n^{(m')}=2,\Xi^{(m')}=d,F_n^{(m')}=a\right)=\frac{1}{\mathbb{Z}}\bar{x}_n(d)\rho(d)\binom{d}{a}\alpha_n^d(1-\alpha_n)^{(d-a)}f_2(d,a) \tag{4.14}$$

where $\mathbb{Z}$ is a normalization function representing the total fitness in the network:

$$\begin{aligned}\mathbb{Z}=&\sum_{d=1}^{D}\sum_{a=1}^{d}\bar{x}_n(d)\rho(d)\binom{d}{a}\alpha_n^d(1-\alpha_n)^{(d-a)}f_2(d,a)\\+&\sum_{d=1}^{D}\sum_{a=1}^{d}(1-\bar{x}_n(d))\rho(d)\binom{d}{a}\alpha_n^d(1-\alpha_n)^{(d-a)}f_1(d,a).\end{aligned} \tag{4.15}$$

Second, given that m' is chosen to reproduce, it selects a neighbour uniformly at random to influence. The node m' has a cooperating neighbours and d defecting neighbours, so a defecting neighbour is influenced with probability:

$$\bar{P}_{21}(d,a,\alpha_n)=\mathbb{P}\left(s_{n+1}^{(m)}=1|s_n^{(m)}=2,s_n^{(m')}=2,\Xi^{(m')}=d,F_n^{(m')}=a,\alpha_n\right)=\frac{d-a}{d} \tag{4.16}$$

We also present a third updating method, death-Rebirth updating, because the population dynamics can be expressed using the SIS dynamics presented in Sec. 2.2.

1. death-Rebirth updating (Diffusion): One individual in the population is selected uniformly at random to die. This focal individual is removed and a new individual is born in its place. The newborn then chooses a strategy according to the fitness of each

strategy available to it, given its neighbourhood.

$$\bar{P}_{21}(d,a,\alpha_n) = \mathbb{P}\left(s_{n+1}^{(m)} = 1 | s_n^{(m)} = 2, \Xi^{(m)} = d, F_n^{(m)} = a, \alpha_n\right) = \frac{f_1^\dagger(d,a)}{f_1^\dagger(d,a) + f_2^\dagger(d,a)}. \quad (4.17)$$

4.2.2 Mean Field and Pair Approximations to the Moran Process

Here we present mean field state evolutions for each of the updating scenarios described in Sec II.

Mean Field Dynamics: death-Birth

In the death-Birth mean field approximation, a node m of degree d and state $s_n^{(m)} = 2$ (defecting strategy) is chosen at time n with probability $\rho d(1-\bar{x}_n(d))$, and this node has a neighbours with a cooperating strategy with probability $\binom{d}{a}\alpha_n^a(1-\alpha_n)^{(d-a)}$, finally a cooperating neighbour is chosen to influence m with probability given in 4.13. Analogous probabilities exist for calculating the probability that a cooperating strategy is replaced by a defecting strategy. Thus the update equation for population state $\bar{x}_n$ in the death-Birth scenario is:

$$\begin{aligned} \bar{x}_{n+1}(d) - \bar{x}_n(d) &= \frac{1}{M}\left(\mathcal{T}^+(\bar{x}_n,d) - \mathcal{T}^-(\bar{x}_n,d)\right) \\ \mathcal{T}^+(\bar{x}_n,d) &= \sum_{a=0}^{d}\binom{d}{a}\alpha_n^a(1-\alpha_n)^{(d-a)}\frac{af_1(d,\alpha_n)}{af_1(d,\alpha_n)+(d-a)f_2(d,\alpha_n)}(1-\bar{x}_n(d)) \\ \mathcal{T}^-(\bar{x}_n,d) &= \sum_{d'=0}^{d}\binom{d}{a}\alpha_n^a(1-\alpha_n)^{(d-a)}\frac{(d-a)f_2(d,\alpha_n)}{af_1(d,\alpha_n)+(d-a)f_2(d,\alpha_n)}\bar{x}_n(d) \end{aligned} \quad (4.18)$$

Recall that α_n is a function of $\bar{x}_n$. In the Moran process, when the prisoner's dilemma is played under weak selection $w << 1$ there may paradoxically be a surviving fraction of cooperators in the network.

Mean Field Dynamics: Birth-death

In the Birth-death mean field approximation, a node is chosen with probability proportional to its fitness, so a node with a cooperating strategy and degree d is chosen and replaces a defecting node m with probability.

$$\begin{aligned}\mathbb{P}\left(s_n^{(m)}=1|s_n^{(m)}=2\right)=&\mathbb{P}\left(s_n^{(m)}=1|s_n^{(m)}=2,s_n^{(m')}=2,\Xi^{(m')}=d,F_n^{(m')}=a\right)\\&\times\mathbb{P}\left(s_n^{(m')}=2,\Xi^{(m')}=d,F_n^{(m')}=a\right)\\&=\frac{1}{\mathbb{Z}}\sum_{d=1}^{D}\sum_{a=1}^{d}\bar{x}_n(d)\rho(d)\binom{d}{a}\alpha_n^d(1-\alpha_n)^{(d-a)}f_2(d,a)\frac{d-a}{d}\end{aligned}$$

For convenience we define

$$\begin{aligned}\Omega_1&=\frac{1}{\mathbb{Z}}\sum_{d=1}^{D}\sum_{a=1}^{d}(1-\bar{x}_n(d))\rho(d)\binom{d}{a}\alpha_n^d(1-\alpha_n)^{(d-a)}f_1(d,a)\frac{a}{d}\\\Omega_2&=\frac{1}{\mathbb{Z}}\sum_{d=1}^{D}\sum_{a=1}^{d}\bar{x}_n(d)\rho(d)\binom{d}{a}\alpha_n^d(1-\alpha_n)^{(d-a)}f_2(d,a)\frac{d-a}{d}\end{aligned}$$

If a cooperating node reproduces and replaces a defecting node then the probability that the defecting node has degree d is given by $\frac{\rho(d)(1-\bar{x}_n(d))d}{\sum_d\rho(d)(1-\bar{x}_n(d))d}$. We notice, however, that the focal node can have any degree and still reproduce on a node of different degree. Thus we must sum the contributions from all degrees and our update is given:

$$\begin{aligned}\bar{x}_{n+1}(d)-\bar{x}_n(d)&=\frac{1}{M}\left(\mathcal{T}^+(\bar{x}_n,d)-\mathcal{T}^-(\bar{x}_n,d)\right)\\\mathcal{T}^+(\bar{x}_n,d)&=\frac{\rho(d)(1-\bar{x}_n(d))d}{\sum_d\rho(d)(1-\bar{x}_n(d))d}\frac{\Omega_2}{\mathbb{Z}}\\\mathcal{T}^-(\bar{x}_n,d)&=\frac{\rho(d)\bar{x}_n(d)d}{\sum_d\rho(d)\bar{x}_n(d))d}\frac{\Omega_1}{\mathbb{Z}}\end{aligned}\tag{4.19}$$

All degrees have had their dynamics tied together by the terms $\frac{\Omega_i}{\mathbb{Z}}, i \in \{1,2\}$. This mean field approximation does not allow for nodes of different degree to reach different equilibria. The reason why cooperation cannot persist when the prisoner's dilemma is played is that, due to the nature of random replacement (regardless of fitness), local structures do not promote cooperation.

Mean Field Dynamics: death-Rebirth (Diffusion)

In the death-Rebirth mean field approximation, a node m of degree d and state $s_n^{(m)} = 2$ (defecting strategy) is chosen at time n with probability $\rho d(1 - \bar{x}_n(d))$, and this node has a neighbours with a cooperating strategy with probability $\binom{d}{a}\alpha_n^a(1-\alpha_n)^{(d-a)}$. It now changes its state according to its own fitness in its neighbourhood. It therefore becomes a cooperator with probability proportional to the fitness of the cooperating strategy. Analogous probabilities exist for calculating the probability that a cooperating strategy is replaced by a defecting strategy. Thus the update equation for population state $\bar{x}_n$ in the death-Rebirth scenario is:

$$\begin{aligned}
\bar{x}_{n+1}(d) - \bar{x}_n(d) &= \frac{1}{M}\left(\mathcal{T}^+(\bar{x}_n, d) - \mathcal{T}^-(\bar{x}_n, d)\right) \\
\mathcal{T}^+(\bar{x}_n, d) &= \sum_{a=0}^{d}\binom{d}{a}\alpha_n^a(1-\alpha_n)^{(d-a)}\frac{f_1^\dagger(d,a)}{f_1^\dagger(d,a)+f_2^\dagger(d,a)}(1-\bar{x}_n(d)) \\
\mathcal{T}^-(\bar{x}_n, d) &= \sum_{d'=0}^{d}\binom{d}{a}\alpha_n^a(1-\alpha_n)^{(d-a)}\frac{f_2^\dagger(d,a)}{f_1^\dagger(d,a)+f_2^\dagger(d,a)}\bar{x}_n(d)
\end{aligned} \tag{4.20}$$

Because the fitnesses of the reproducing node does not depend on the current cooperation state of the network, the dynamics of the moran process with death-Rebirth updating can be represented exactly with the SIS model. The death-Rebirth Moran process can be tracked using the optimal Bayesian filter of Chapter 3.

Mean field approximations do not incorporate local structural information that can be effect long term cooperation (or infection) levels. We can extend our analysis to determine the effect network structure has on long term cooperation. This can be done using the pair approximation.

Pair Approximation: death-Birth

The mean field description of spreading cooperation does not incorporate local structural information. The network structure is assumed to be entirely encoded in the degree distribution, and the population is assumed to be mixed such that there are no connection correlations between nodes of the same state. That is, cooperating nodes do not have a higher probability of being connected to other cooperating nodes than defecting nodes being connected to cooperating nodes. This assumption can be relaxed using the pair approximation. Local information can be critical to appropriately characterizing the Moran process. Under weak selection prisoner's dilemma cannot result in a positive steady state fraction of cooperators, however under the pair approximation, the steady state cooperating population state may have a positive fraction of cooperators. This is because local areas of cooperators may be more fit than defectors, resulting in higher population fitness, even when the individual stategy for defecting is always the better (dominant) individual choice. In the pair approximation, the state of the network is summarized not simply as the degree segregated cooperation probabilities, $\bar{x}_n$, but also the conditional probabilities that a neighbour of a cooperating individual is cooperating/defecting, the conditional infected population state $\bar{u}$, where:

$$\bar{u}_{i|j}(d) = \frac{1}{M}\sum_{m=1}^{M} \mathcal{I}\big(\Xi^{(m)} = d, s_n^{(m)} = i \big| s_n^{(g)} = j : g \in V : g, m \in E\big), \quad d = 1, \dots, D. \quad (4.21)$$

These conditional probabilities provide a more local description of the cooperation state. Note that we can also represent this conditional probability as the joint probability that a node of degree d and a randomly chosen neighbour are both cooperators $\bar{u}_{11}(d)$ over the probability that a node of degree d is a cooperator $\bar{x}_n(d)$, that is $\bar{u}_{i|i}(d) = \frac{\bar{u}_{ii}(d)}{\bar{x}_n(d)}$. We can represent the state of the network can be described in as an augmented state:

$$\bar{z}_n(d) = [\bar{x}_n(d), \bar{u}_{11}(d)], \quad d = 1, \ldots, D. \tag{4.22}$$

We can express all joint and conditional probabilities relating cooperating and defecting neighbour pairs as follows:

$$\bar{u}_{11}(d) = \bar{u}_{1|1}(d) \cdot \bar{x}_n(d) \tag{4.23}$$

$$\bar{u}_{21}(d) = \bar{u}_{12}(d) = (1 - \bar{u}_{1|1}(d))\bar{x}_n(d) \tag{4.24}$$

$$\bar{u}_{22}(d) = 1 - \bar{x}_n(d)(2 - \bar{u}_{1|1}(d)) \tag{4.25}$$

$$\bar{u}_{2|2}(d) = 1 - \frac{\bar{x}_n(d)(1 - \bar{u}_{1|1}(d))}{(1 - \bar{x}_n(d))} \tag{4.26}$$

$$\bar{u}_{1|2}(d) = \frac{\bar{x}_n(d)(1 - \bar{u}_{1|1}(d))}{(1 - \bar{x}_n(d))} \tag{4.27}$$

$$\bar{u}_{2|1}(d) = 1 - \bar{u}_{1|1}(d) \tag{4.28}$$

Pair approximation can, for instance, more accurately describe a population in which there are high cooperator-cooperator/defector-defector correlations. These correlations may allow cooperation to persist. The superiority of this approximation is made evident in the setting where two dense network substructures are connected by a chain of individuals, such that the network may be visualized in a dumbbell shape. Consider network dynamics wherein a node m has fitness 1 if it has more cooperating neighbours than defecting neighbours:

$$\bar{P}_{21}(d, a, \alpha_n) = \begin{cases} 1, & a > \frac{d}{2} \\ 0, & a < \frac{d}{2} \\ 0.5, & a = \frac{d}{2} \end{cases}, \quad \bar{P}_{12}(d, a, \alpha_n) = \begin{cases} 1, & a < \frac{d}{2} \\ 0, & a > \frac{d}{2} \\ 0.5, & a = \frac{d}{2} \end{cases}$$

If one dense subnetwork is of all cooperators and the other all defectors, then a positive fraction of cooperators is expected in the steady state system. Consider this same scenario under the mean field approximation, here the populations of cooperators and defectors are presumed mixed and the steady state system will drift towards the state (cooperating/defecting) of the largest subnetwork. Clearly, the pair approximation provides a better description of the system dynamics. In evolutionary game theory, these local state correlations are important as it is known that in regular networks, the pair approximation supports a positive fraction of cooperators in the defection dominated prisoner's dilemma under death-Birth updating. Previous derivations of heterogeneous pair approximation did not lend themselves to the Moran process with general fitness, motivating us to explore further [64] [13].

Much like in the mean field approximation we have derived degree dependent update equation, only now we must update both the cooperating population state and the conditional cooperating population state simultaneously. To perform this update, we generalize the cooperating link probability α_n to the conditional cooperating link probabilities $\alpha_{n,i|j}$ for states $i, j \in \{1, 2\}$:

$$\begin{aligned}\alpha_{i|1}(\bar{z}_n) &= \frac{\sum_{d=1}^{D} (\text{\# of links from infected node of degree } d \text{ and type 1 to type i})}{\sum_{d=1}^{D} (\text{\# of links of degree } d \text{ and type 1})} \\ &= \frac{\sum_{d=1}^{D} d\,\rho(d)\,\bar{u}_{i|1}(d)}{\sum_{d=1}^{D} d\,\rho(d)\bar{x}_n(d)}.\end{aligned} \tag{4.29}$$

With this conditional link probability, we can represent the expected payoff of a cooperating/defecting individual:

$$\pi_1(d, \alpha_{n,1|1}, \alpha_{n,2|1}) = \mathbf{a}{\cdot}\alpha_{n,1|1}d{+}(d{-}\alpha_{n,2|1}d)\mathbf{b} \qquad \pi_2(d, \alpha_{n,2|2}, \alpha_{n,1|2}) = \mathbf{c}{\cdot}\alpha_{n,1|2}d{+}(d{-}\alpha_{n,2|2}d)\mathbf{d}. \tag{4.30}$$

In the death-Birth pair approximation, the infected population state changes in much the same way as in the mean field approximation, only now the cooperating link probability depends on the state of the node being chosen. A node m of degree d and state $s_n^{(m)} = 2$

(defecting strategy) is chosen at time n with probability $\rho d(1-\bar{x}_n(d))$, and this node has a neighbours with a cooperating strategy with probability $\binom{d}{a}\alpha^a_{n,1|2}(1-\alpha_{n,1|2})^{(d-a)}$, finally a cooperating neighbour is chosen to influence m with probability given in 4.13. Analogous probabilities exist for calculating the probability that a cooperating strategy is replaced by a defecting strategy. Thus the update equation for population state $\bar{x}_n$ in the death-Birth scenario is:

$$\begin{aligned}
\bar{x}_{n+1}(d)-\bar{x}_n(d) &= \frac{1}{M}\left(\mathcal{T}^+(\bar{x}_n,\bar{u}_{11},d)-\mathcal{T}^-(\bar{x}_n,\bar{u}_{11},d)\right)\\
\mathcal{T}^+(\bar{x}_n,\bar{u}_{11},d) &= \sum_{a=0}^{d}\binom{d}{a}\alpha^a_{n,1|1}(1-\alpha_{n,1|1})^{(d-a)}\frac{af_1(d,\alpha_n)}{af_1(d,\alpha_n)+(d-a)f_2(d,\alpha_n)}(1-\bar{x}_n(d))\\
&= \sum_{a=0}^{d}\mathcal{U}^+(\bar{x}_n,\bar{u}_{11},d,a)\\
\mathcal{T}^-(\bar{x}_n,\bar{u}_{11},d) &= \sum_{d'=0}^{d}\binom{d}{a}\alpha^a_{n,1|2}(1-\alpha_{n,1|2})^{(d-a)}\frac{(d-a)f_2(d,\alpha_n)}{af_1(d,\alpha_n)+(d-a)f_2(d,\alpha_n)}\bar{x}_n(d)\\
&= \sum_{a=0}^{d}\mathcal{U}^-(\bar{x}_n,\bar{u}_{11},d,a)
\end{aligned} \tag{4.31}$$

Where $\mathcal{U}^+(\bar{x}_n,\bar{u}_{11},d,a)$ is the contribution of nodes of degree d with a cooperating neighbours to the increase in the cooperating population state. For each individual m that changes its state from cooperating to defecting the joint probability of an edge connecting two cooperating individuals decreases by $\frac{a}{d}$, and there is a similar increase for defecting to cooperating transitions. We can, therefore, update the joint probability of neighbouring cooperating nodes as:

$$\begin{aligned}
\bar{u}_{11,n+1}(d)-\bar{u}_{11,n}(d) =& \frac{1}{M}\sum_{a=0}^{d}\left(\frac{a}{d}\left(\mathcal{U}^+(\bar{x}_n,\bar{u}_{11},d,a)-\mathcal{U}^-(\bar{x}_n,\bar{u}_{11},d,a)\right)\right)\\
&+\sum_{d'=0,d'\neq d}^{D}\sum_{a=0}^{d'}\left(\frac{a\rho(d)}{d}\left(\mathcal{U}^+(\bar{x}_n,\bar{u}_{11},d',a)-\mathcal{U}^-(\bar{x}_n,\bar{u}_{11},d',a)\right)\right)
\end{aligned} \tag{4.32}$$

The second term accounts for the probability that a change in a node not of degree d causes a change in the probability that a cooperating node of degree d interacts with another cooperating node of arbitrary degree.

4.2.3 Tracking Cooperation in Game theoretic Networks

In Chapter 3, we presented a non-linear optimal filter for tracking the infected population state of a network. This filter cannot be directly applied to the cooperation models presented in 4.2.2 for all updating methods because the mean field approximation and pair approximations presented in these sections do not have a polynomial state evolution. In the mean field approximation, the in death-Birth and Birth-death updating, the update contains a term in the denominator which depends on the state[18]. In the pair approximation, no matter which updating type is chosen (although we have only explored death-Birth in this thesis), the state does not have a polynomial evolution. As with the simpler polynomial infected population state, we are motivated to track the cooperating population state. Through the sampling mechanisms presented in Sec. 2.3, we can generate observations of the cooperating population state. The Bayesian polynomial filter of Chapter 3 cannot be applied without approximation, and so we are motivated to utilize other filtering methods. We, therefore, utilize the suboptimal filtering method, particle filtering, to tracked the cooperating population state.

In this thesis, we provide a brief overview of particle filtering for the nonlinear dynamics of the presented mean field and pair approximations to the Moran process, for a more complete exposition we refer readers to [7]. Particle filtering is a Monte Carlo algorithm for recursive Bayesian estimation of the system state, here the system state is the cooperating population state. In particle filtering, weighted random samples are used to approximate the posterior densities of the state prediction and update. As the number of particles increases, this

[18] We can however note that death-Rebirth updating has polynomial dynamics that match the polynomial dynamics of Sec. 2.1.3.

approximation converges to the true posterior densities and the approximation approaches an optimal Bayesian estimate.

As in the polynomial filter, the filtering problem consists of prediction and updating. The posterior density is approximated by a set of N_p particles and weights $\{x_n^i, w_n^i\}^{N_p}$:

$$p(x_n|y_n) \approx \sum_i w_n^i \delta(x_n - x_n^i)$$

At each time step we sample particles from a proposal density $x_n^i \sim \pi(x_{0:n}, Y_n)$ and each particle is weighted according to

$$w_n^i = \frac{p(x_{0:n}^i|y_{1:n})}{\pi(x_{0:n}^i|Y_n)}$$

and we can approximate the filtered posterior density as:

$$p(x_n|y_n) \approx \sum_i \frac{w_n^i}{\sum_j w_n^j} \delta(x_{0:n}^i) \tag{4.33}$$

The posterior state estimate can be computed directly from the filtered posterior density, and, therefore, approximated by the weighted particles:

$$\hat{x}_{n+1} = \mathbb{E}(x_n|Y_n) \approx \sum_i x_n^i \frac{w_n^i}{\sum_j w_n^j}$$

The proposal density is updated in real time

$$\pi(x_{0:n}|Y_n) = \pi(x_0) \prod_{t=1}^{n} \pi(x_n|x_{0:n-1}, Y_n) \tag{4.34}$$

$$\pi(x_n|x_{n-1}, Y_n) = p(x_n|x_{n-1}^i, y_n) = \frac{p(y_n|x_n)p(x_n|x_{n-1})}{p(y_n|x_{n-1})}$$

which in turn yields a recursion for the particle weights

$$w_n^i = \frac{p(x_{0:n}^i|Y_n)}{\pi(x_{0:n}^i|Y_n)} \propto \frac{p(y_n, Y_{n-1}, x_{0:n}^i)p(x_n^i|x_{0:n-1}^i)}{\pi(x_n^i|x_{0:n-1}^i, Y_n)} w_{n-1}^i. \tag{4.35}$$

From Eq. 4.35 and Eq. 4.35, we obtain the optimal update equation for the particle weights.

$$w_n^i = p(y_n|x_{n-1}^i)w_{n-1}^i \tag{4.36}$$

Since our system has nonlinear dynamics and linear measurements 2.19 and both have Gaussian noise with known covariance, there is a closed form expression $p(y_n|x_{n-1})$ and $p(x_n|x_{n-1}, y_n)$. First we define mean vector m and precision matrix Σ. Recall that Q denotes the state noise covariance and R denotes the observation noise covariance. From eq. 4.18, we can express the mean field evolution of the state, x_n, with death-Birth updating as the nonlinear function $f(x)$:

$$x_{n+1}(d) = f(x_n) = x_n(d) + \frac{1}{M}\left(\mathcal{T}^+(x_n, d) - \mathcal{T}^-(x_n, d)\right).$$

The state update can be expressed as similar nonlinear functions for the pair approximation as well as mean field approximations with other updating types. Utilizing this functional update as well as the state and observation noise covariances we define

$$\begin{aligned}\Sigma^{-1} =& \mathbf{Q}^{-1} + C'\mathbf{R}^{-1}C \\ m =& \Sigma\left(\mathbf{Q}^{-1}f(x_{n-1}) + C'\mathbf{R}^{-1}y_n\right)\end{aligned} \tag{4.37}$$

Utilizing this mean vector and covariance we present analytic evaluations of the conditional densities which can be used to re -sample particles and update their respective weights:

$$\begin{aligned}p(y_n|x_{n-1}) =& \mathcal{N}(Cf(x_{n-1}), (\mathbf{Q} + C\mathbf{R}C')) \\ p(x_n|x_{n-1}, y_n) =& \mathcal{N}(m, \Sigma).\end{aligned} \tag{4.38}$$

The presented particle filtering algorithm does not depend on the state evolution having polynomial dynamics and can, therefore, be employed to filter the mean field death-Birth and

Birth-death Moran processes, as well as the pair approximation to these Moran processes. This more general filtering method is not optimal for finite numbers of particles and, as such, the state estimates produced using particle filtering are inferior to the state estimates of the polynomial Bayesian filter presented in Chapter 3 when the underlying state has polynomial dynamics. Even when the underlying system follows polynomial dynamics, it may be sensible to employ a particle filter over the Bayesian filter. The particle filter computation time scales with the number of particles, whereas the polynomial filter computation time scales with the dimension of the system. Since the dimension of the proposed system scales directly with the maximum degrees of the considered networks, for extremely large networks, it may be sensible to utilize a particle filtering method to generate state estimates.

4.3 Summary

This chapter presented two extensions to the SIS model of chapter 2. The first extension considered was the addition of homophily in the network, wherein individuals were delineated by type, and the frequency and outcome of interactions between individuals depended upon the types of individuals involved. The second extension considered is an alternative underlying model, the Moran process. The Moran process is a model for network evolutionary dynamics which is considered in the context of modeling game theoretic cooperation. The dynamics of the Moran process are approximated using both the mean field approximation considered in chapter 2 and the pair approximation. A particle filter is presented for tracking the cooperating population state of a network.

Chapter 5

Numerical Results

Sec. 5.1 first presents an example of an infected population state, observations of the infected population state, and a filtered state estimate. The parameters for this network are determined empirically from the Twitter dataset, as outlined explicitly in Sec. 5.4.

Sec. 5.2 below examines the effect of sampling and model misspecification on the performance of the non-linear filter discussed in Sec.3.1. This analysis is useful for selecting the sampling methodology and for assessing the performance trade-off due to imperfect knowledge of the degree distribution of the underlying network.

Sec. 5.3 presents numerical examples of filtering on networks with homophily as well as filtering on game theoretic networks. Additional network simulations illustrate the role of homophily in modeling SIS dynamics and mean field approximations to the Moran process.

Sec. 5.4 presents the performance analysis on a real-world Twitter dataset. First, using a goodness-of-fit test, we validate the sufficiency of the SIS model to capture the infection propagation on the Twitter network. Second, the non-linear filter of Sec.3.1 is shown to track the infection diffusion satisfactorily.

5.1 Example: Tracking the Infected Population State

Following *Scheme B* of Sec. 3.2, we generated scale-free networks with $M = 13000$ (number of nodes) to emulate the Twitter Social network. The infections were initialized by infecting nodes at time 0 with probability 0.01. This infection was propagated for 2×10^4 timesteps[19] with a healing probability $\xi = 0.001$ and $\bar{P}_{12}$ generated empirically from Twitter data of

[19] Recall from Theorem 1 that the duration is of the order of number of nodes.

Sec. 5.4. At each timestep, 10^4 samples were obtained according to the uniform sampling described in Sec.2.3. The state and observation covariance matrices $\mathbf{Q}_n$ and $\mathbf{R}_n$ used in filtering were computed empirically from the true underlying state and observation data, and were estimated to be constant for all time steps n. The resulting observations and filtered estimates are shown in Fig. 5.1, where the network was simulated, sampled and filtered according to 'Scheme B' of Sec. 3.2. It can be seen that the estimates converge towards the true state for all degrees. The mean square error of the filter estimates are shown in Fig. 5.12 and compared to the filtered estimate for the Twitter data of Sec. 5.4. The displayed mean square errors are the average of 50 independent simulations.

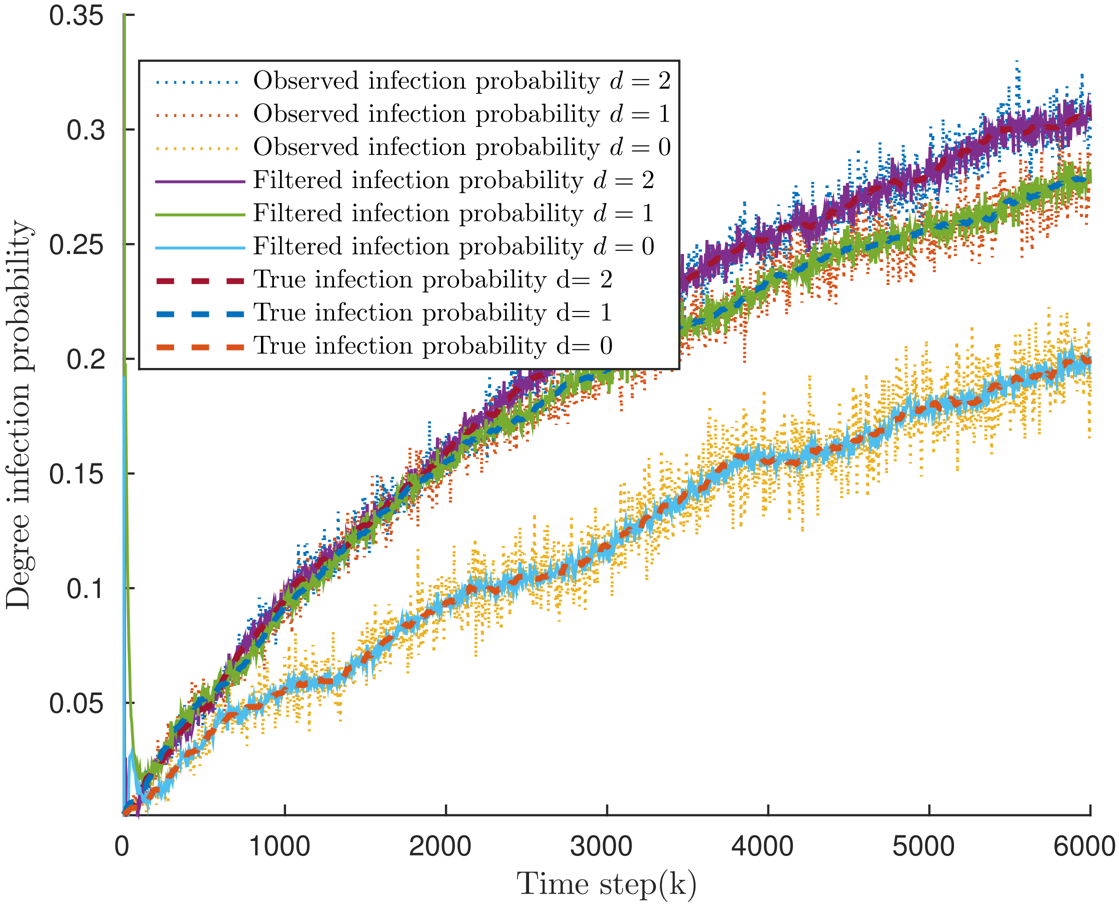


Figure 5.1: Diffusion of infected population states and their corresponding filtered estimates in a scale-free network.

5.2 Factors Affecting Filter Performance

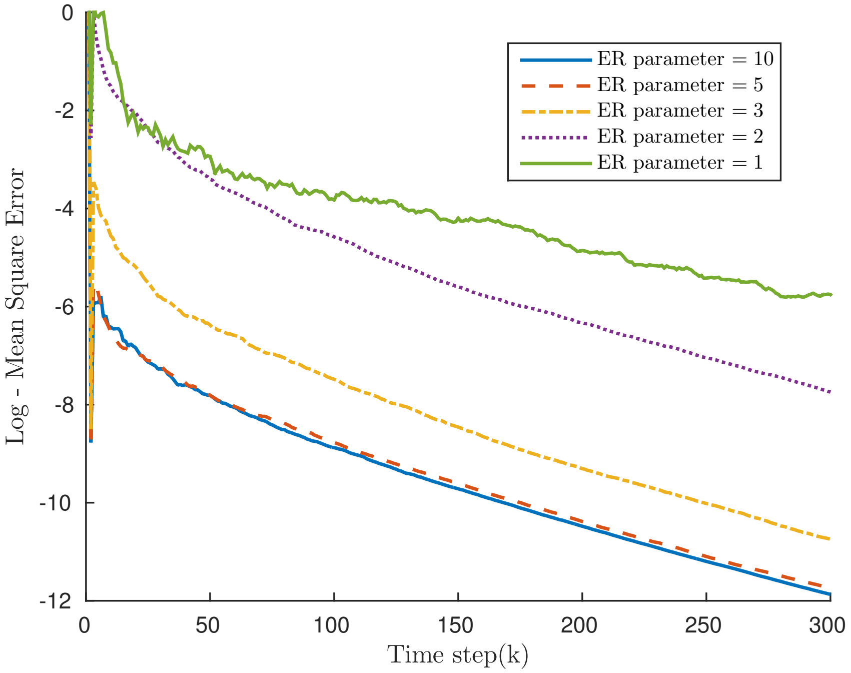
As in Sec. 3.2, we consider ER and SF networks and analyze the effect of sampling and model misspecification on the filter performance. To isolate these simulation tests from the approximation of $\mathbf{R}_n$, as well as state noise, we utilize Scheme A of Sec. 3.2, which simulates a deterministic trajectory with noisy observations.

5.2.1 Effect of Sampling on Filtering

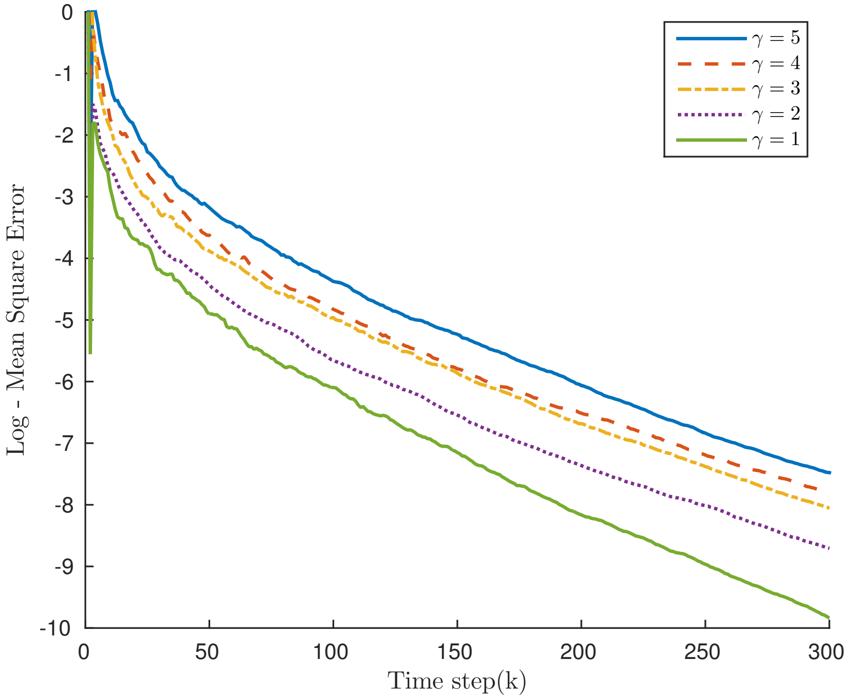
The observation noise variance $\mathbf{R}$ in (2.19) depends on the sampling method employed. Using the uniform sampling mechanism outlined in Sec.2.3, the effect of sampling on filtering is illustrated.

Under uniform sampling, the noise variance of each observation depends upon the network degree distribution. In the uniform sampling of Sec. 2.3, for any given degree, an observation at time n will have a variance which corresponds to that of a scaled binomial distribution $\sigma^2(\hat{x}_n(d)) = \frac{x_n(d)(1-x_n(d))}{\nu(d)}$, where $\nu(d)$ are the number of samples of nodes of degree d and $\nu = \sum_d \nu(d)$. For a large number of independent samples, by the central limit theorem, $\nu(d) \approx \nu\rho(d)$, and the observation noise variance is inversely related to the probability that a node is of degree d. Consequently, the noise covariance depends upon degree distribution parameters γ and λ. In Fig. 5.2, the mean square error of the filter estimate is seen to depend on the network parameters λ and γ.

In these simulations, the observation error covariance matrices were chosen as the error covariance matrices from *network based simulations*(Scheme B). We isolate the experiment from finite network state noise, which can dominate the MSE of the filter estimate. By simulating the system with Scheme A, with observation noise informed by the synthetic network simulations of Scheme B, we can observe the effect of sampling more clearly.



(a) Filtered estimate log mean square error of infection population states on Erdős Rényi degree distributions.



(b) Filtered estimate log mean square error of infection population states on scale-free degree distributions.

Figure 5.2: Filtered estimate log mean square error of infection population states for two network types: Erdős Rényi and scale-free.

5.2.2 Sensitivity of Filter Performance to Misspecified Model

The SIS model in Sec. 2.1.2 is misspecified if the degree distribution ρ is not known/specified exactly. The Bayesian filter (Sec. 3.1) implemented using a misspecified model is referred to as a misspecified filter. This section compares the MSE of a Bayesian filter and a misspecified filter, both formulated for the same underlying network. The misspecified filter was derived with the degree distribution $\rho(d) = \frac{1}{D} \quad \forall d$. For further comparison, we use a moving average (linear) estimator:

$$\hat{\hat{x}}_n = \vartheta_1 y_{n-1} + \vartheta_2 y_{n-2} + \vartheta_3 y_{n-3} + \cdots + \vartheta_\iota y_{n-\iota} + \vartheta_0. \tag{5.1}$$

Here $\hat{\hat{x}}_n$ is the moving average estimate at time n, the matrices ϑ_i are computed using multivariate least squares estimation, and time delay ι was chosen to be 10, for simplicity. In Fig. 5.3, the infection transition probabilities $\bar{P}_{12}$ and $\bar{P}_{21}$ were chosen uniformly at random over $[0, 1]$. The observation noise covariance matrix R is a random positive definite matrix with entries in $[0, 3 \times 10^{-2}]$. The initial state was simulated as $x_0 = \frac{1}{2} \forall d$. The misspecified filter exhibits a plateau in performance. This plateau corresponds to the incorrect computation of the priori state estimate, $\hat{x}_n^-$, due to the misspecified filter parameters. It is observed in Fig. 5.3 that, even when the degree distribution is misspecified, the Bayesian filters outperform the moving average filter with an MSE of the order of 10^{-6}, compared to 10^{-4} of the moving average filter.

5.3 Extended and Alternative Diffusion Models

5.3.1 Numerical Exploration of Homophily

In this section, we illustrate the performance of the filter on a social network having homophily as an additional attribute. The network was simulated according to Scheme B of Sec. 3.2 with two population types, b and c, as shown in Fig. 5.4. These populations interact

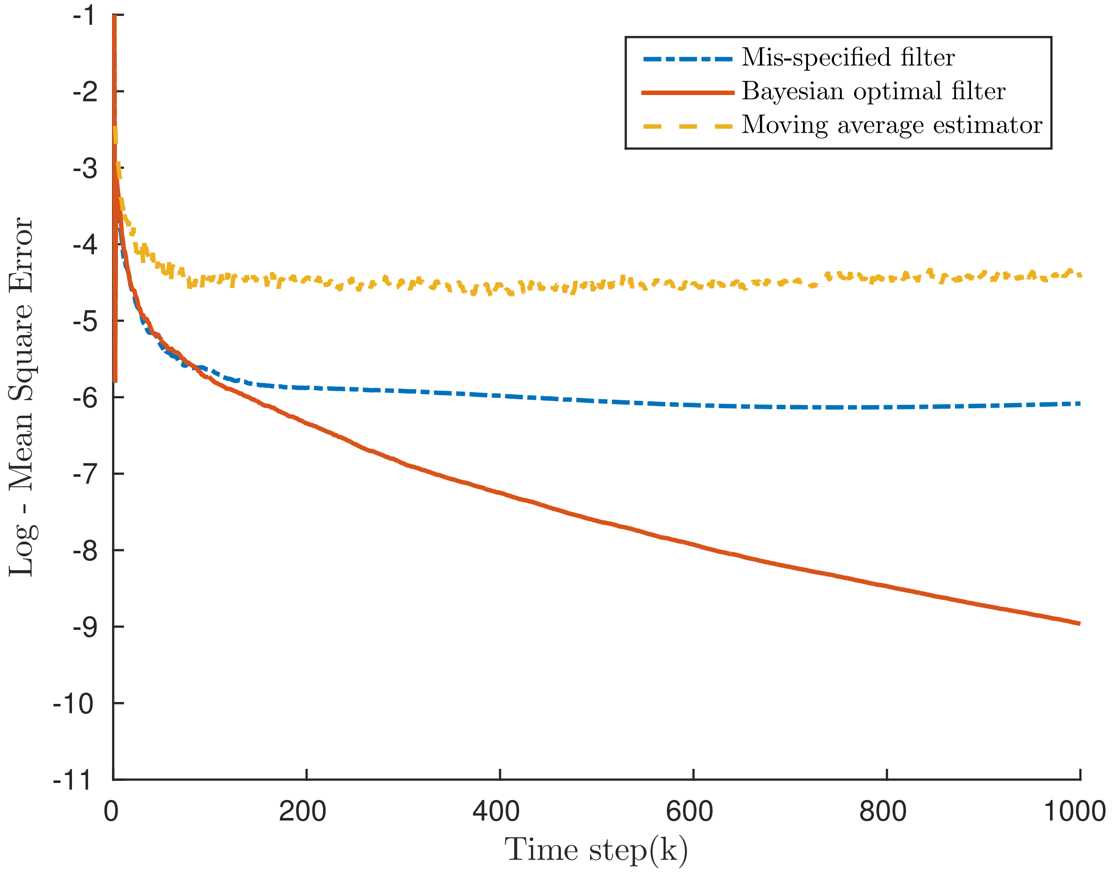


Figure 5.3: Comparison of the estimate log mean square error between a Bayesian filter, a misspecified Bayesian filter and a moving average estimator.

more with members of their own type than those of the other type, specifically $\varsigma_{bb} = \varsigma_{cc} = 0.8$. To differentiate the populations, we choose the infection probabilities for population b to be four times[20] that of population c , that is $\bar{P}^b_{21} = 4\bar{P}^c_{21}$. The recovery probabilities and scale-free degree distributions are the same for both populations. In Fig. 5.4, we plot the infected population state, observed state and filtered state estimate for both population types, using degree $d = 2$ as an example. The network was simulated according to 'Scheme B' of Sec. 3.2 and each node was labeled as type b or c with equal probability. The network was then sampled and filtered according to 'Scheme B' of Sec. 3.2. The filtered estimate of the infected population state converges to the true infected population state. The two different

[20] Population b is therefore more resistant to infection.

population types can be observed to behave differently as the low infection probabilities of the type c population and insular network structure, cause a low infected population state in nodes of c.

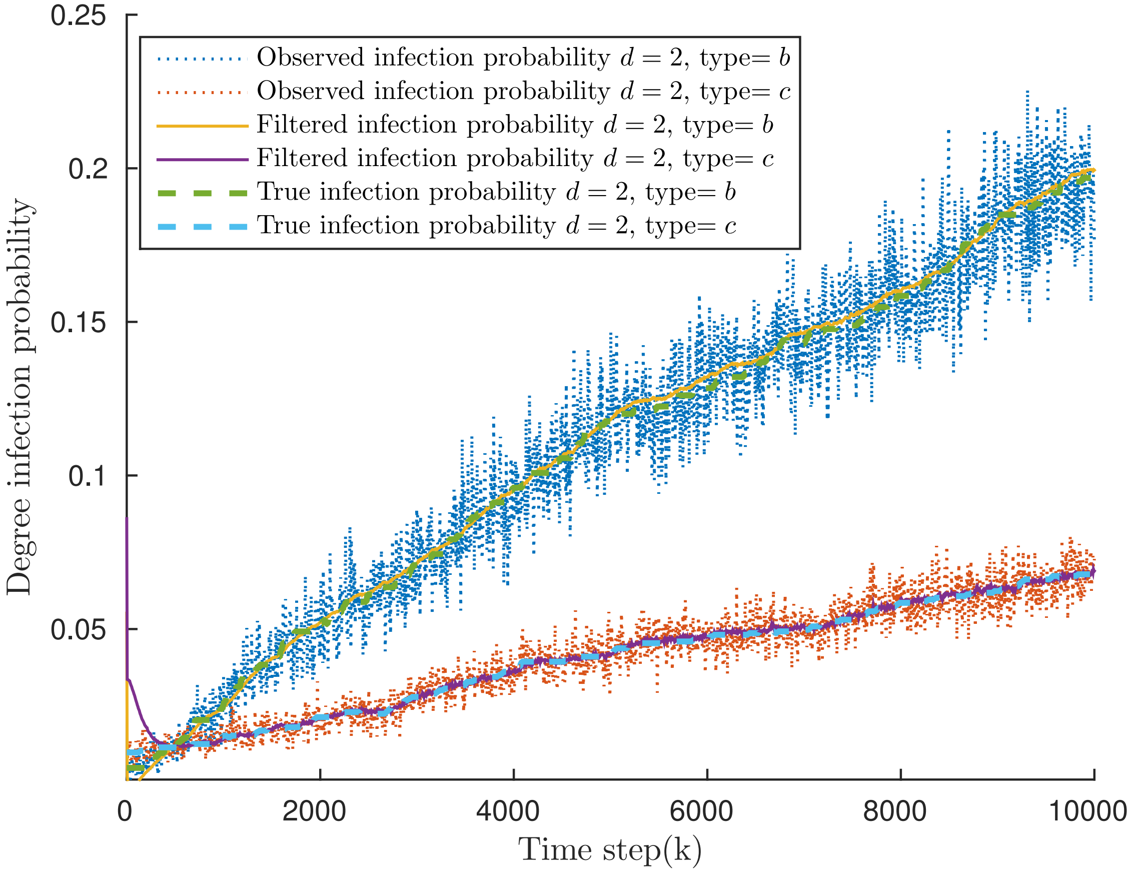


Figure 5.4: Diffusion of infected population states and their corresponding filtered estimates in a scale-free network.

The SIS model with homophily also enables the use of additional degree distribution information to generate a more accurate mean field description of a system. In a network the degree distribution of nodes attached to each degree d may not equal the degree distribution of the entire network. For example, solitary people may be more likely to also associate with other solitary people, resulting in high level of internal connections between all nodes of low degree. Thus the degree distribution of low degree nodes may differ from the degree distribution of high degree nodes. These degree-degree correlations are not accounted for in the SIS model presented in Sec. 2.2, however, we can use the SIS model with homophily to incorporate this information. To model additional degree information, we map each degree in the network to a type b, so that the number of types $\mathcal{B}$ equals the maximum degree in the network plus one[21] $D+1$, and allowing each type to have its own degree distribution $\rho^{[b]}(d)$, which may be different from the network's degree distribution ρ.

In Fig. 5.5, we illustrate the effect additional degree distribution information can have on the dynamics of a network by simulating and comparing the deterministic mean field trajectories for two SIS models, one with homophily and one without. We consider a scale-free network segregated by degree, where individuals with d neighbours associate primarily with others having d neighbours. The network has an overall degree distribution $\rho(d) \propto d^{-\gamma}$, where $\gamma = 3.0$, and maximum degree $D = 20$. The degree distribution of each type $b \in \{1, \dots, \mathcal{B}\}$ is given by:

$$\rho^{[b]}(d) = (1-\epsilon)\mathcal{I}(d = b-1) + \epsilon\rho(d). \tag{5.2}$$

where $\mathcal{I}$ is an indicator function. In Fig. 5.5, the infection probabilities, maximum degree, and overall degree distributions are identical for both systems. The degree distributions used in both simulations are scale-free degree distributions, ρ, with $\gamma = 3.0$, and the degree specific degree distributions, $\rho^{[b]}(d)$, in the simulation with homophily is given by (5.2) with $\epsilon = 0.02$. The additional degree distribution information incorporated into the model with homophily changes the system dynamics, and a difference between the models with and without homophily can be observed. The key observation here is that under a simple SIS model as presented in Sec. 2.2, both these systems are identical, however, we can utilize additional degree distribution information and the SIS model with homophily to better capture the mean field dynamics of the population.

[21] There are $D+1$ types because degree $d = 0$ is also modeled.

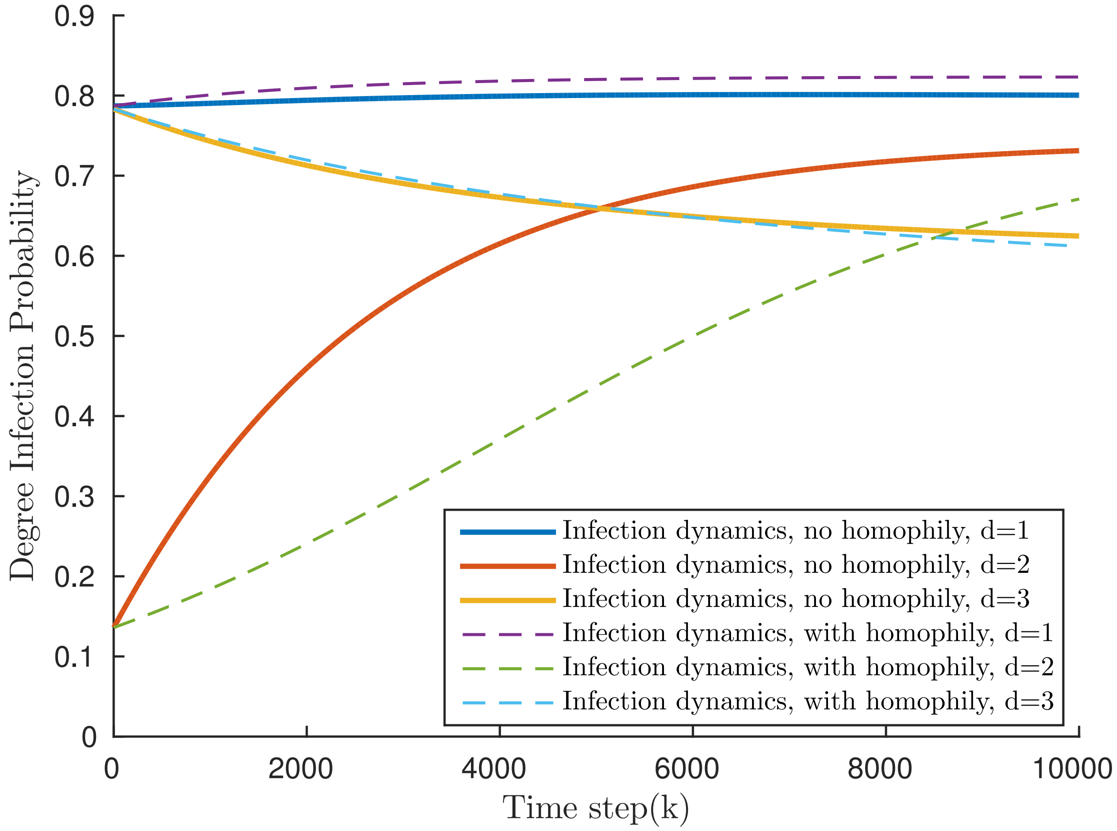


Figure 5.5: Deterministic mean field trajectories of an SIS model with and without homophily are compared.

5.3.2 Numerical Exploration of Game Theoretic Dynamics

In this section, we investigate game theoretic network dynamics. The networks structures investigated were generated with $N = 10,000$ and were Erdős Rényi networks with parameter $\lambda = 2.0$. We observed the dynamics of the prisoner's dilemma with payoffs $[a = 4, b = 0, c = 5, d = 2]$ and death-Birth updating. In Fig. 5.6, we plot the cooperating population state, observed state and filtered state of degree $d = 2$ as an example. The filtered state estimate is generated by the particle filter described in Sec. 4.2.3 using 1000 particles and is seen to track the true cooperation state. In Fig. 5.7 and Fig. 5.8, we plot the mean field state and simulated state under strong ($w = 0.9$) and weak selection ($w = 0.01$), respectively. The mean field approximation is fairly accurate under strong selection, however, it fails to accurately describe the population state under weak selection. Under weak selection, the cooperating population state does not decay towards 0 as predicted by mean field dynamics. Instead, local structures can promote cooperation as predicted in regular networks [56].

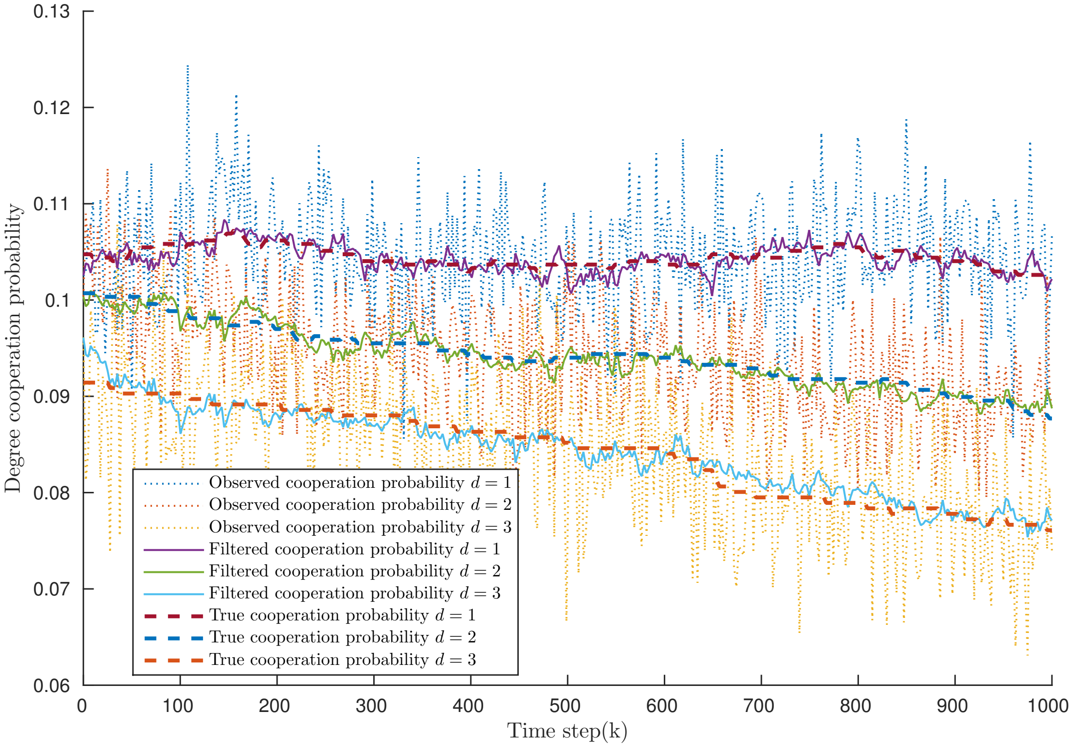


Figure 5.6: Cooperating population state of an Erdős Rényi network, as well as state observations, and particle filter state estimate.

5.4 Analysis and Validation on Twitter Dataset

This section illustrates the tracking of infection diffusion on social networks using real diffusion data from the microblog platform Twitter. Twitter is a social media platform over which users can communicate in short < 140 character messages, images and video files. We analyze the diffusion of information through the Twitter Social network to demonstrate the effectiveness of the SIS model of Sec. 2.2.

Twitter played a critical role in the Egyptian revolution of 2011 or January 25^{th} (#Jan25)

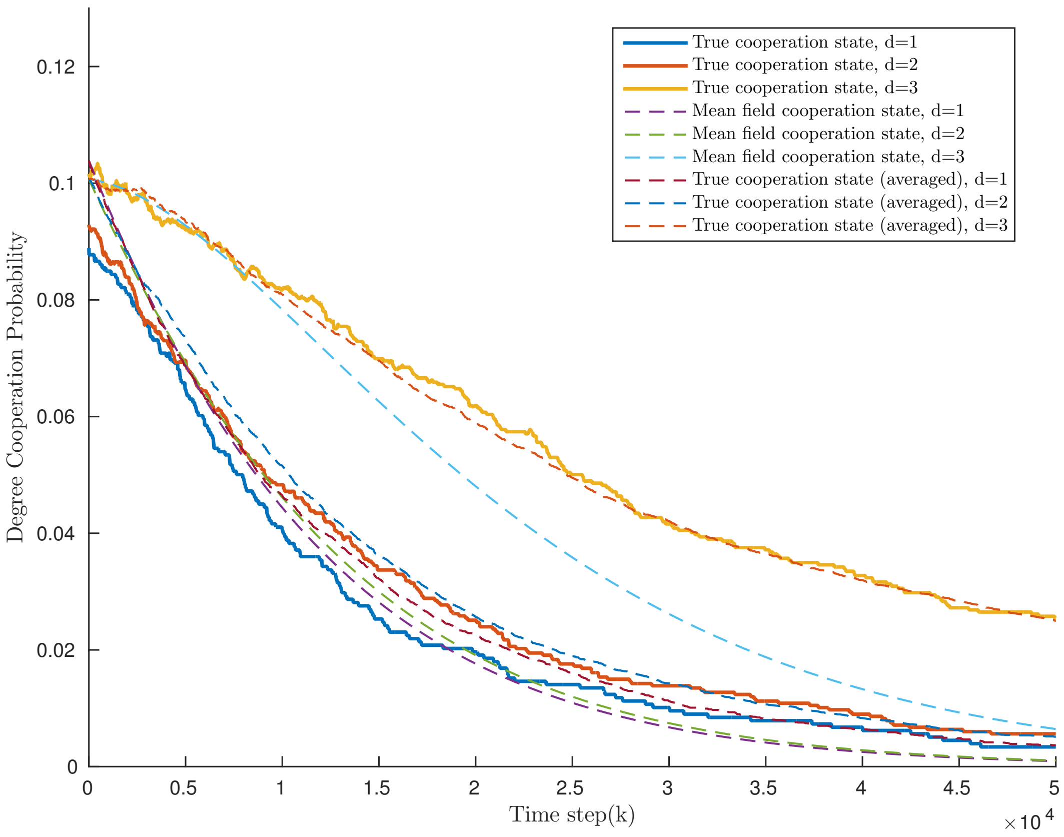


Figure 5.7: Simulated and mean field approximation of cooperating population state under strong selection $w = 0.9$.

uprising [70]. Twitter was used by protesters to organize the protest and recruit members and as a medium to discuss and share information about the protest. Below, we refer to the interest and engagement with the news of the uprising as *infection* and track the distribution of infection over time. Modeling this online process as an epidemic is supported by the virality of the #Jan25 hashtag. Viral online information is found to follow complex contagion models in [84].

5.4.1 Dataset

The dataset consists of tweets sampled between January 23^{rd} and February 8^{th}, 2011 and are available from Twitter (http://trec.nist.gov/data/tweets/). The tweet collection period

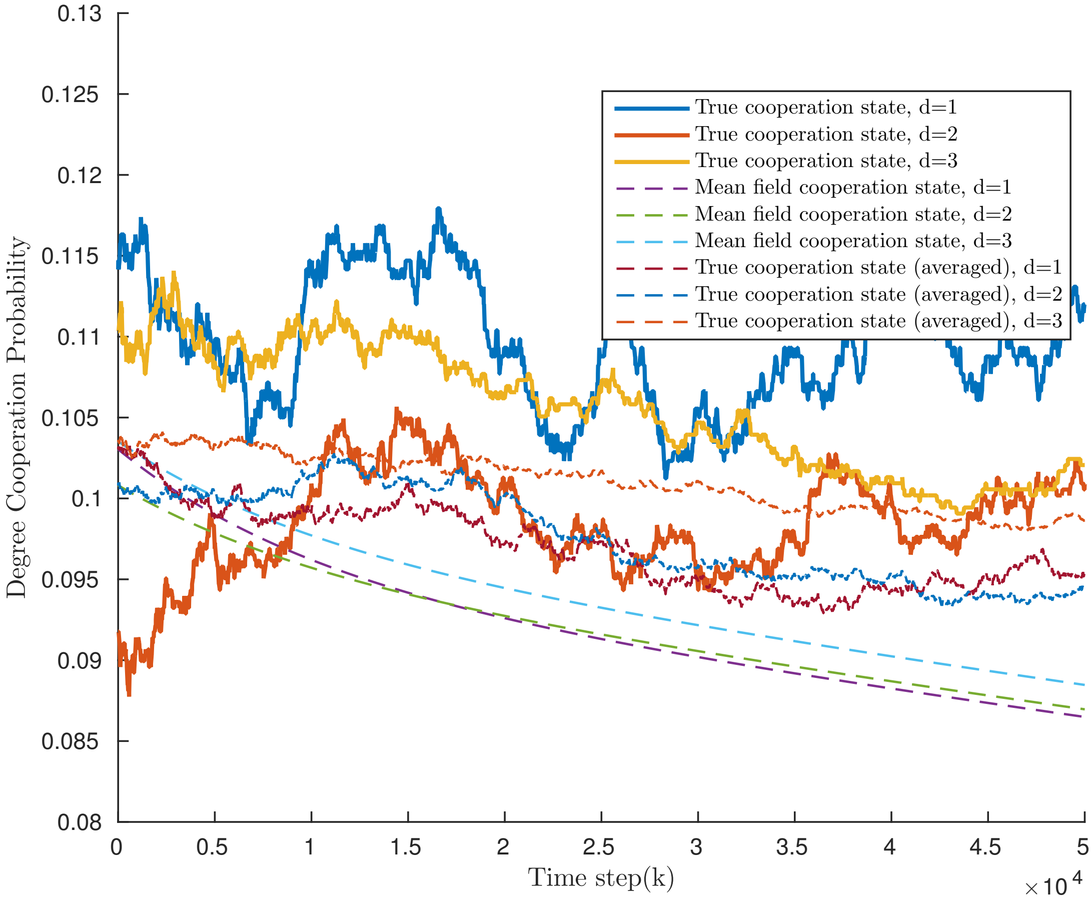


Figure 5.8: Simulated and mean field approximation of cooperating population state under weak selection $w = 0.01$.

encapsulates the time-frame of the first major developments relating to the January 25^{th} uprising event. In Twitter, a "hashtag" follows the discussion topics, i.e., a word or a phrase prefixed with the number sign #. We make use of the hashtags to track the spreading of a specific topic on Twitter. The most used hashtag related to this protest is "#Jan25". To obtain the information spreading among users participating in this protest, we filtered 26,313 tweets containing "#Jan25" published by 13,341 different users, from around 10 million tweets. These tweets contain the event of interest and the social network is (re-)constructed from them as follows: two users are connected if one user has mentioned another user ("@username") in the tweet containing "#Jan25" at least once over the duration of interest.

On this constructed social network, information diffusion is analysed.

All users in the constructed social network are assumed to be susceptible initially. Users who initiate tweets on the event of interest are assumed to be infected and act as seeds for the spread of information. Once a user, say User#A becomes infected, it has some constant probability, say ξ, of becoming susceptible in each time period. This modeling assumption is motivated by the frequently observed Poisson-like decay of an individual's interest in social media topics [23]. The decay probability was chosen to be 0.001, motivated heuristically by an average interest duration of 2 hours. Our dataset is thus a hybrid dataset, wherein the network and infection process is informed by true Tweets between users, and these infected individuals become susceptible according the synthetic Poisson-like process described above. The effect the choice of ξ has on the infection dynamics can be seen in Fig. 5.9. The decay parameter does affect the twitter infection trajectory, however, for decay parameters that correspond to an average forgetting timescale on the order of hours, the trajectories appear similar.

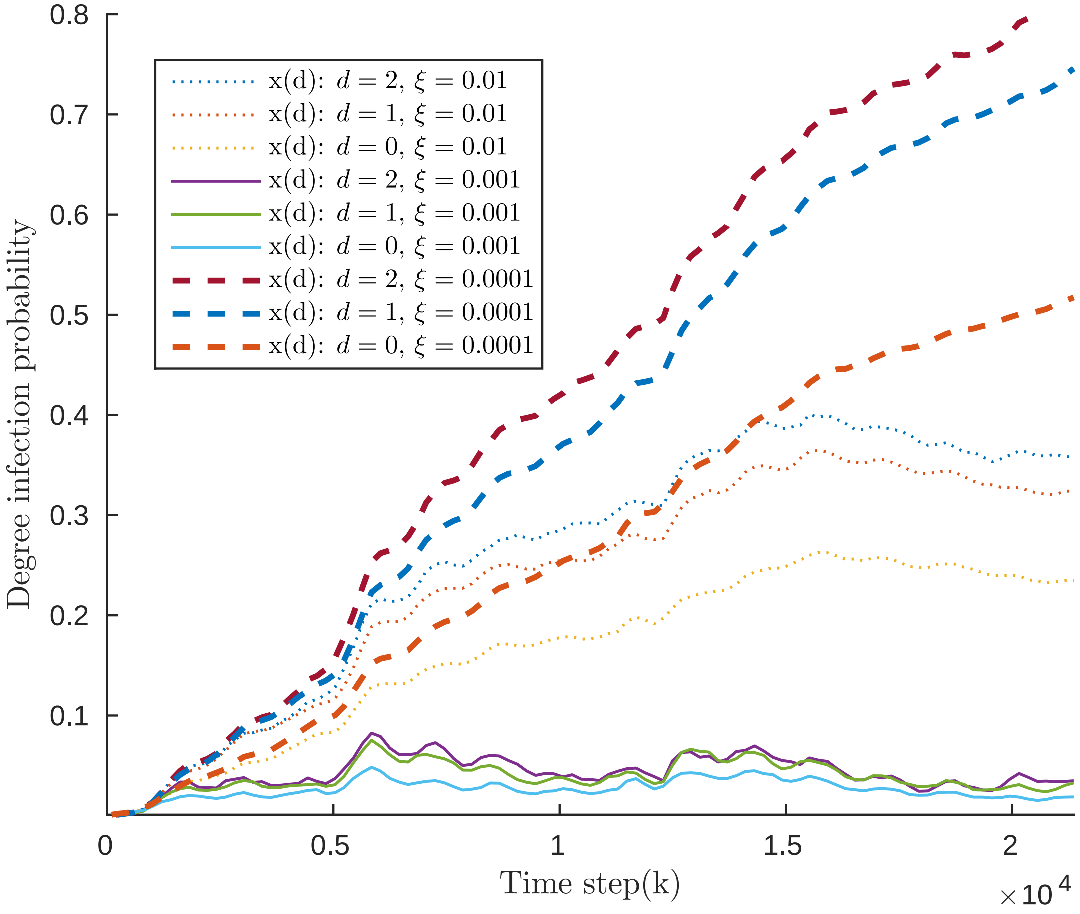


Figure 5.9: Comparison of infected population state of $d = 1, 2, 3$ for varying decay parameter ξ.

5.4.2 SIS model for Twitter data

In case of the spread of engagement in Twitter, individuals can either be inactive or active depending on if they are willing and able to spread information and interest on a topic. Active users can be considered 'infected', and inactive users can become 'infected' by interacting with other 'infected' individuals, in particular, any of its active neighbors in a social network. In this way, engagement and knowledge of a topic spreads throughout the network. People can also become disinterested in a subject they have already been exposed to, in this way they are not currently engaged, but may become engaged if contacted by an infected neighbor;

thus inactive individuals are assumed to be susceptible. The SIS model has been applied to information diffusion in [51], [46].

We adopt the SIS model for the Twitter data, mapping engagement in the January 25^{th}, 2011 uprising as an infection spreading over the Twitter network created from Twitter users. Below we analyze the goodness-of-fit of the SIS model for the Twitter data using the Kolmogorov-Smirnov test [24], and measure the square and absolute difference between the data and model predictions.

Model Evaluation (Goodness of Fit for SIS model): We used the KS test on the empirical infected degree distribution at the final timepoint to evaluate the goodness of fit of the SIS model. The KS test statistic was 0.2286 with p-value 0.2813. The null hypothesis for this statistical test is that both the observed Twitter and SIS model infected degree distributions are samples of the same infected degree distribution. At a confidence level of 0.01, the null hypothesis cannot be rejected. This test therefore provides little evidence for or against the quality of SIS model for Twitter. To further explore the agreement between the model and the true data, we computed the average and maximum square difference between the Twitter data and predicted SIS degree infection probabilities. These values can be seen in Table 5.1 and the trajectories are shown in Fig. 5.10. The underlying degree distribution of the Twitter network is scale-free, and thus degree and number of nodes of that degree are inversely related. In Fig. 5.10, the simulated trajectories use $\xi = 0.001$ and the empirically generated ρ and $\hat{\hat{P}}_{12}$ (5.3). Recall that without added state noise, or finite networks, the simulations for the mean field dynamics are deterministic. The deterministic mean field dynamics effectively capture the trajectory of the Twitter infection. Empirically, we observe that the fewer nodes there are of a particular degree, the higher the probability that the empirical infection deviates from the asymptotic case.

The low magnitude of the model deviations in Table 5.1 for the Twitter dataset and the failure to reject the hypothesis that the Twitter data and model infected degree distributions come from the same distribution, suggest that the SIS model is a satisfactory model with

Table 5.1: Goodness of fit for the SIS model: The average and maximum deviations between the Twitter data and SIS model predictions are presented.

Degree	1	2	3+
Average Square Difference	0.0011	0.0014	0.0235
Average Absolute Difference	0.0273	0.0294	0.1000
Maximum Absolute Difference	0.0644	0.0719	0.8403

respect to the infection dynamics of interest in the January 25th uprising.

Sampling for tracking the infected population state

The mean field dynamics for the SIS model can be used to track and predict the evolution of the infection on Twitter. We must generate estimates of $\bar{P}_{12}$, $\bar{P}_{21}$, and determine the degree distribution from samples obtained from (2.19). $\bar{P}_{12}$ is given by ξ, since all infected nodes become susceptible with probability ξ at each time point. We compute the empirical transmission rates $\hat{\bar{P}}_{21}$ directly by observing the frequency with which an infected individual with d neighbors, a of which are infected, becomes infected.

$$\hat{\bar{P}}_{12}(d,a) = \frac{\sum_{n=0}^{T}\sum_{m=1}^{M} \mathcal{I}\left(s_{n+1}^{(m)} = 2 | s_n^{(m)} = 1, \Xi^{(m)} = d, F_n^{(m)} = a\right)}{\sum_{n=0}^{T}\sum_{m=1}^{M} \mathcal{I}\left(s_n^{(m)} = 1, \Xi^{(m)} = d, F_n^{(m)} = a\right)} \tag{5.3}$$

The degree distribution used is the empirical degree distribution. The true degree infection probabilities are computed directly from the entire network for each 1 minute time interval.

Next, we sample the data using the RDS sampling scheme described in Sec. 2.3, every 1 minute and track the infected population states over time using the non-linear Bayesian Filtering technique described in Sec. 3.1. The parameters used in the Bayesian filter are $\xi = 0.001$, the empirically generated ρ and $\hat{\bar{P}}_{12}$. The filter estimates are shown in Fig. 5.11. It

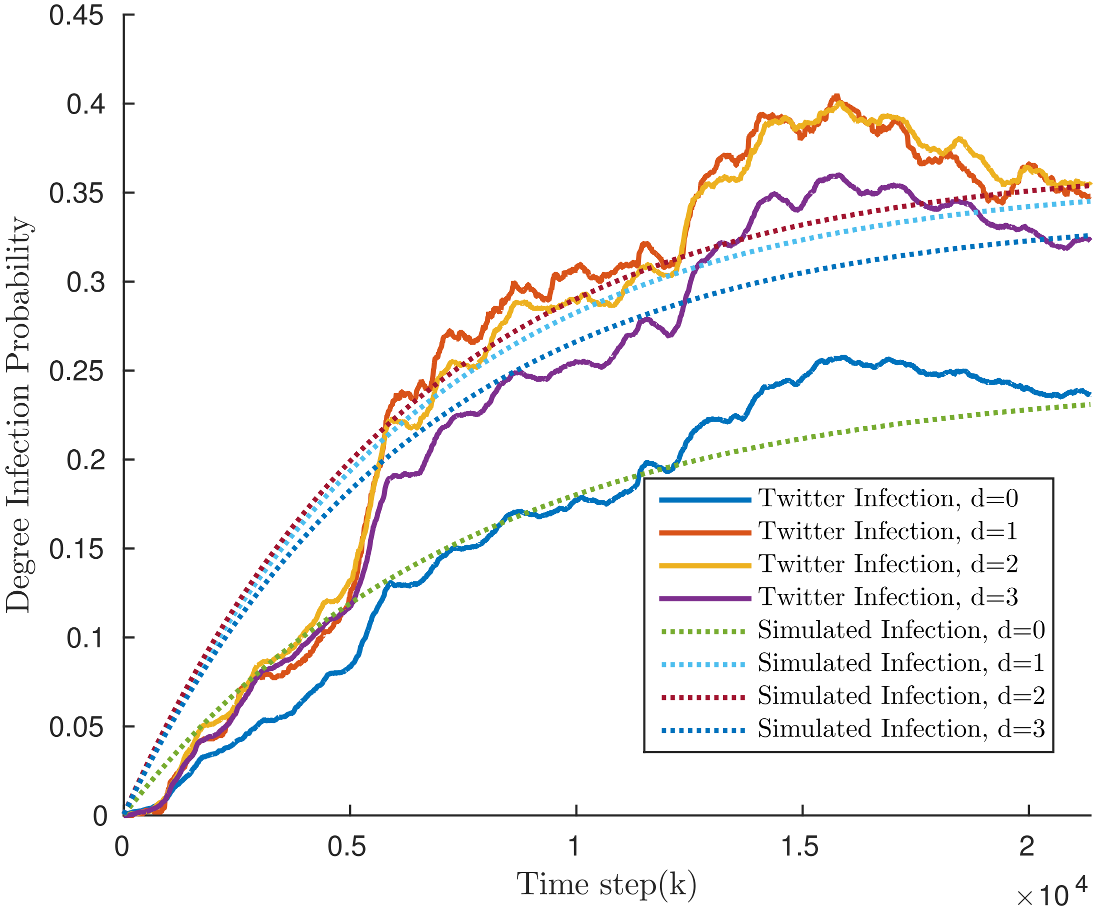


Figure 5.10: The true Twitter infection and Twitter infection as predicted by the deterministic mean field dynamical equations are compared for nodes of degree $d = 1, 2, 3$.

is seen that the filtered estimates track the true infected population states over time. Fig. 5.12 shows the mean square errors of the filter estimate[22] of Twitter infection and the mean square errors of the filter estimate of the numerical example shown in Fig. 5.1. This figure illustrates the superiority of the filter on the numerical example and highlights the shortcomings of the Twitter dataset. The filter performs adequately for both the Twitter data and the simulated network, however the estimate for the simulated network, the infection of which follows the SIS model, performs dramatically better at a filtered estimate MSE of 10^{-11} versus the filtered estimate of the Twitter data around 10^{-6}.

5.5 Summary

In this Chapter, we numerically illustrated the performance of the filtering algorithms presented in Chapters 3 and 4 on scale-free and Erdős Rényi networks. We investigated the sensitivity of the non-linear filtering algorithm to errors in model specification. Further, the model with homohphily was numerically investigated and estimates of simulated infected population states were generated using the non-linear filtering algorithm of Chapter 3, and the game theoretic model based on the Moran process was investigated and filtered using the particle filter presented in 4. We then introduced, motivated and analyzed a real world dataset from Twitter. This dataset consists of a corpus of tweets taken during the Arab Spring uprising of 2011. The non-linear filtering algorithm was applied to this dataset and shown to satisfactorily estimate the infected population state.

[22] The MSE of the Twitter estimate starts extremely low. During early timesteps in the Twitter data, there are almost no infected nodes. The sampled observations similarly estimate nearly 0 infection, which is why the filtered infection appears so accurate at these early timesteps. Thus, as the system dynamics evolves from the trivial 0 infection, the state observations become less accurate.

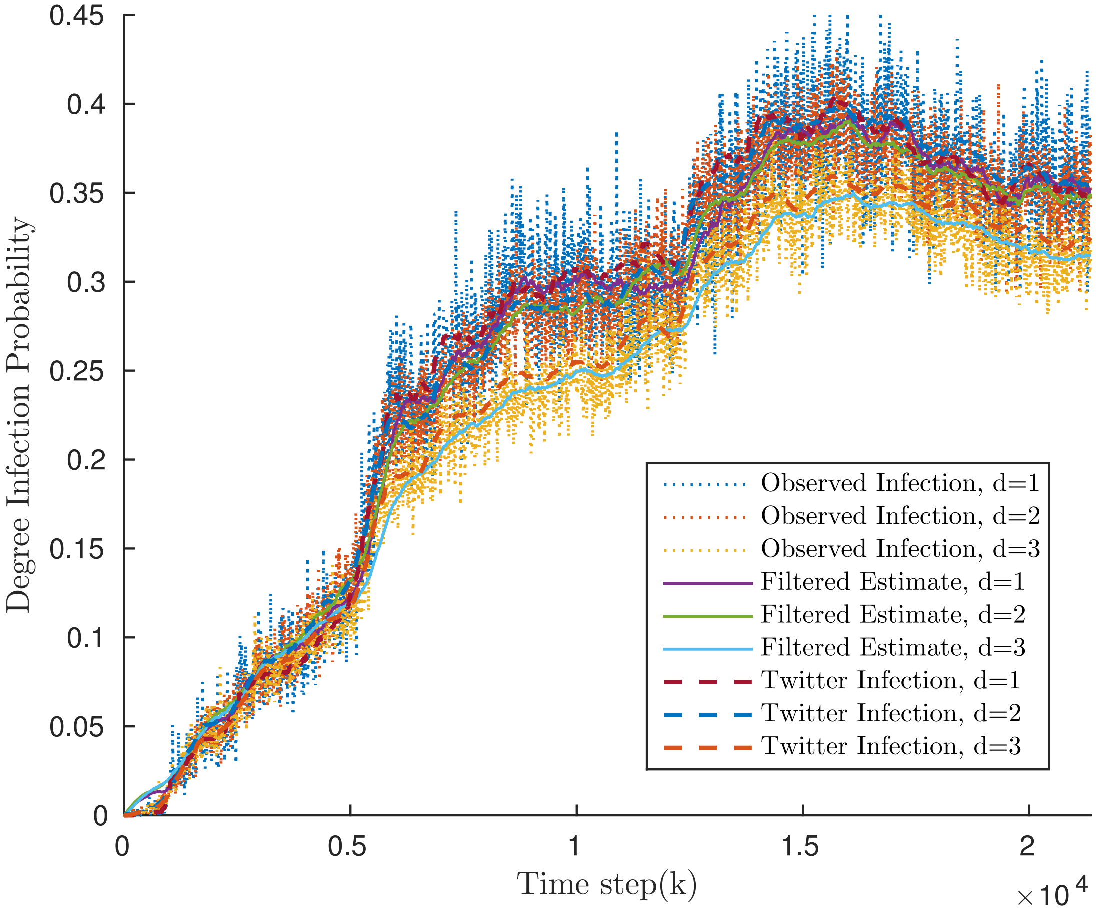


Figure 5.11: The true Twitter infection and the filter estimates of the Twitter infected population state are compared.

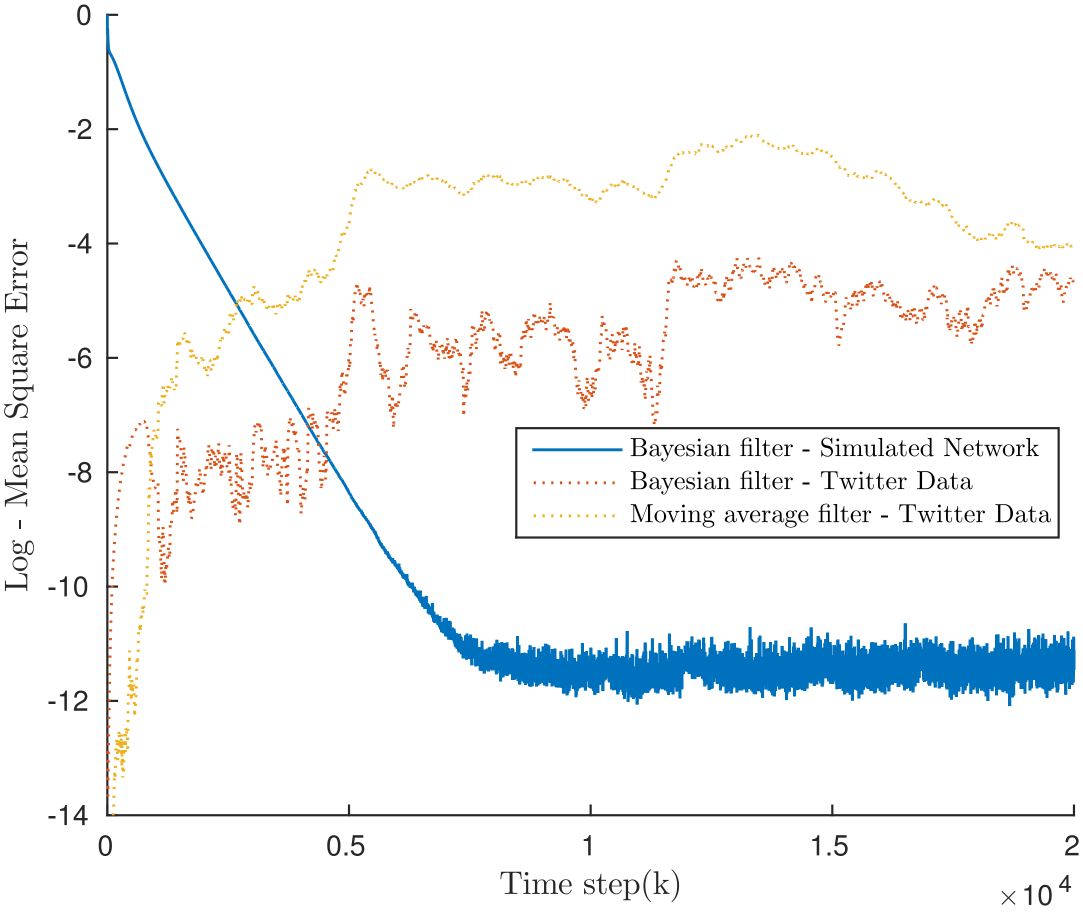


Figure 5.12: The MSE of the filtered estimate of the true twitter infection probability state and the MSE of the estimate of the infected population state of a simulated network are shown.

Chapter 6

Conclusions and Future Work

In this thesis, we developed a method for tracking infections, information and opinions as they diffuse through social networks. By utilizing statistical signal processing filtering algorithms, we are able to combine domain knowledge of the underlying infection process and observations to generate accurate estimates of the infected population state. Tracking diffusion in social networks is applicable to the commercialization, political analysis, and fundamental understanding of large online social networks as well as policy making in public health through deepening our epidemiological understanding of the evolution of a disease.

The second chapter presented a formulation of infection diffusion dynamics over social networks, a mean field approximation to this complex process, and finally sampling methods for generating observations of the underlying infected population state. Two types of social network structures were primarily considered: Erdős-Rényi and scale-free networks. An infection diffusion model was developed based on the SIS model. We presented a general description of mean field theory and its ability to approximate complex interacting systems. Mean field dynamics for the SIS model were then formally presented, including a theorem which states that the difference between the underlying population state and the mean field approximation satisfies an exponential bound, and therefore the mean field dynamics can be utilized as a model which can generate the observable infection dynamics. Under discrete time mean field dynamics, the state at time $n+1$ depends polynomially on the state at time n. These polynomial dynamics are exploited by the optimal Bayesian filter of Chapter 3.

The third chapter describes an optimal non-linear Bayesian filter for estimating the infected population state. First the polynomial mean field dynamics are presented so that the polynomial dependence is explicit. A non-linear Bayesian filter is then described for optimally estimating the infected population state of the social network. The fundamental limits of filtering the mean field dynamics are then explored through the computation of the PCRLB. The PCRLB is computed for Erdős-Rényi and scale-free networks.

In the fourth chapter we incorporate homophily into to the SIS model of chapter 2 and present a game theoretic contagion model. The homophily extension improves the generality and expressive power of the underlying SIS model. The model with homophily allows the population to be stratified by type, permits these different types to interact with different frequencies, and enables the use of different infection transmission probabilities based on type. The model with homophily does this without sacrificing any of the polynomial properties of the state evolution that make the model amenable to optimal Bayesian filtering. We also present a game theoretic contagion model which uses the Moran process as the underlying model for infection dynamics. This model, which is similar to the SIS model, allows for interactions to be game theoretic. Under the mean field and pair approximations we develop in this chapter, this Moran model introduces a dependence of the infection transmission probabilities on the infected/cooperating population state of the network. Finally, in this chapter we present a particle filter for tracking the infected/cooperating population state under when the underlying infection follows this Moran model.

The fifth chapter presented numerical results substantiating the infection diffusion model and filtering algorithms presented in earlier chapters. In this section, first, the effect of network structure on the filtered infected population state estimate as investigated. Next, the effect of model misspecification on the mean square error of the filter estimate was investigated. Numerical examples of the influence of homophily as well as examples of filtering game theoretic contagion using a particle filter are presented. The non-linear filtering algorithm is then applied to a dataset based on real world tweets from the social media website Twitter. The filtering algorithm is shown to perform satisfactorily. Finally, a two timescale co-evolving network is simulated and the underlying degree distribution and infected population state are estimated.

In future work, more work could be done to extend the analysis of scale-free and Erdős-Rényi networks. Currently, we have begun to explore the effect that network structure has on our ability to estimate the infected population state on these networks, however, simulations of larger networks may provide useful information to future researchers wishing to track diffusion on networks of known network type. Further, we considered two sampling types in this thesis, RDS sampling and simple random sampling. These sampling methods can be further assessed in the context of large scale-free and Erdős-Rényi networks, with and without homophily, to better understand the effect of the sampling type has on the filter generated state estimates.

In this thesis, the networks and infection parameters we analyzed and investigated were static. The infection transmission probabilities $\bar{P}_{12}/\bar{P}_{21}$ as well as diffusion parameter θ were independent of time n in all numerical investigations, however, the models presented were amenable to dynamics parameters, so long as the dynamics of these parameters is known apriori. The effect of time varying infection parameters can be explored numerically, to determine the effect these parameters have on the mean square error of the filter estimate. A dynamic diffusion parameter is especially relevant to modeling real world interactions, because human contact processes rarely occur at a uniform rate. We may, for instance, tweet more in the morning and evenings, or engage with more people on weekends. Further, future work could be done to extend this work to filtering on evolving networks. In [48], two-timescale network dynamics are outlined, as well as a Hidden-Markov-Model (HMM) filter for tracking evolving network structure. This two-timescale model requires the strong assumption that network dynamics are much slower than infection dynamics, which does not always hold. Analytic and numerical investigations of more complex co-evolving networks could prove enlightening and extend the set of networks on which the infected population state can be adequately filtered.

The game theoretic contagions provide a rich arena for exploration. In this thesis, the prisoner's dilemma was presented as the persistent example of the game underlying the Moran process, but other two-player or even multi-player games can be considered. Games such as the Hawk-Dove game and Stag-hunt game both provide dramatically different cooperation dynamics. Further, the network-based mean field approximations can be analyzed to explore the effect network structure has on the equilibrium states of the various games.

Finally, the filtering model can be applied to more real world data to entrench its efficacy. In this thesis, we analyzed a dataset from Twitter and showed that the polynomial filter performed adequately, however, it was observed that the polynomial filter performed significantly better at filtering the infected population state of synthetic network contagions. This discrepancy may be investigated further both to establish that the real world "infections" are well modeled by the SIS model, as well as utilizing a larger share of the tools at the modeler's disposal, such as dynamic diffusion parameters, infection transmission probabilities, game theoretic contagion models and models with homophily to better model and ultimately filter the infection process.

Bibliography

[1] Facebook, Inc. (2016, Oct. 10). *Company information: Stats* [Online]. Available: https://newsroom.fb.com/company-info/

[2] L. A. Adamic and B. A. Huberman, “Power-law distribution of the world wide web,” *Science*, vol. 287, no. 5461, pp. 2115–2115, 2000.

[3] N. K. Ahmed,J. Neville, and R. Kompellaa, “Network sampling: From static to streaming graphs,” *ACM Transactions on Knowledge Discovery from Data (TKDD)*, vol. 8, no. 2, pp. 7, 2014.

[4] L. Allen, “Some discrete-time SI, SIR, and SIS epidemic models,” *Mathematical Biosciences*, vol. 124, no. 1, pp. 83 - 105, 1994.

[5] J. Alstott, E. Bullmore, and D. Plenz, “powerlaw: a Python package for analysis of heavy-tailed distributions,” *PloS One*, vol. 9, no. 1, pp. e85777, 2014.

[6] T. W. Anderson, “Multivariate statistical analysis,” *Wiley and Sons*, 1984.

[7] M. Arulampalam, S. Maskell, N. Gordon, and T. Clapp “A tutorial on particle filters for online nonlinear/non-Gaussian Bayesian tracking,” *IEEE Transactions on signal processing*, vol. 50, no. 2, pp. 174-188, 2002.

[8] A. Barabási, “Scale-free networks: a decade and beyond,” *Science*, vol. 325, no. 5939, pp. 412–413, 2009.

[9] A. Barabási and R. Albert, “Emergence of scaling in random networks,” *Science*, vol. 286, no. 5439, pp. 509–512, 1999.

[10] A. Barabási, R. Albert, and H. Jeong, “Scale-free characteristics of random networks: The topology of the world-wide web,” *Physica A: Statistical Mechanics and its Applications*, vol. 281, no. 1, pp. 69–77, 2000.

[11] R. Albert and A. Barabási, “Statistical mechanics of complex networks,” *Reviews of Modern Physics*, vol. 74, pp. 47–97, 2002.

[12] M. Benaim and J. Weibull, “Deterministic approximation of stochastic evolution in games,” *Econometrica*, vol. 71, no. 3, pp. 873–903, 2003.

[13] J. Benoit, A. Nunes, and M. Telo da Gama, “Pair approximation models for disease spread,” *The European Physical Journal B-Condensed Matter and Complex Systems*, vol. 50, no. 1, pp. 177–181, 2006.

[14] S. Boucheron, G. Lugosi, and P. Massart, “Concentration inequalities: A nonasymptotic theory of independence,” *Oxford university press*, 2013.

[15] A. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, and J. Wiener, “Graph structure in the web,” *Computer Networks*, vol. 33, no. 1, pp. 309–320, 2000.

[16] P. J. Carrington, J. Scott, and S. Wasserman, “Models and methods in social network analysis,” *Cambridge University Press*, vol. 28, 2005.

[17] C. Castillo-Chavez and A. Yakubu, “Discrete-time SIS models with complex dynamics,” *Nonlinear Analysis: Theory, Methods and Applications*, vol. 47, no. 7, pp. 4753–4762, 2001.

[18] C. Chamley, “Rational Herds: Economic models of social learning,” *Cambridge University Press*, 2004.

[19] N. Chen, “On the approximability of influence in social networks,” *SIAM Journal on Discrete Mathematics*, vol. 23, no. 3, pp. 1400–1415, 2009.

[20] S. Chen, R. Varma, A. Sandryhaila, and J. Kovačević, “Discrete signal processing on graphs: Sampling theory,” *IEEE Transactions on Signal Processing*, vol. 63, no. 24, pp. 6510–6523, 2015.

[21] F. Chung and L. Lu, “Complex Graphs and Networks,” *American Mathematical Society*, vol. 107, 2006.

[22] A. Clauset, C. R. Shalizi, and M. E. Newman, “Power-law distributions in empirical data,” *SIAM Review*, vol. 51, no. 4, pp. 661–703, 2009.

[23] R. Crane and D. Sornette, “Robust dynamic classes revealed by measuring the response function of a social system,” *roceedings of the National Academy of Sciences*, vol. 105, no. 41, pp. 15649–15653, 2008.

[24] D. Darling, “The Kolmogorov-Smirnov, Cramer-von Mises tests,” *The Annals of Mathematical Statistics*, vol. 28, no. 4, pp. 823 –838, 1957.

[25] G. Das, N. Koudas, M. Papagelis, and S. Puttaswamy, “Efficient sampling of information in social networks,” *Proceedings of the ACM workshop on search in Social Media*, pp. 67–74, 2008.

[26] V. Dukic, H.F. Lopes, and N.G. Polson. ”Tracking epidemics with Google flu trends data and a state-space SEIR model.” *Journal of the American Statistical Association*, vol. 107, no. 500, pp. 1410-1426, 2012.

[27] D. Easley and J. Kleinberg, “Networks, crowds, and markets: Reasoning about a highly connected world,” *Cambridge University Press*, 2010.

[28] R. J. Elliott, L. Aggoun, and J. B. Moore, “Hidden Markov models: estimation and control,” *Springer International Publishing*, vol. 29, 2008.

[29] K. Fang and R. Li, “Some methods for generating both an NT-net and the uniform distribution on a Stiefel manifold and their applications,” *Computational Statistics & Data Analysis*, vol. 24, no. 1, pp. 29–46 1997.

[30] O. Frank, “Network sampling and model fitting,” *Models and methods in social network analysis*, pp. 31–56, 2005.

[31] P. Geroski, “Models of technology diffusion,” *Research policy*, vol. 29, no. 4, pp. 603–625, 2000.

[32] G. Ghoshal, L. Chi, and A. L. Barabási, “Uncovering the role of elementary processes in network evolution,” *Scientific Reports*, vol. 3, pp. 2920, 2013.

[33] S. Goel and J. Salganik, “Respondent-driven sampling as Markov chain Monte Carlo,” *Statistics in Medicine*, vol. 28, no. 17, pp. 2202–2229, 2009.

[34] M. Granovetter, “Threshold models of collective behavior,” *American Journal of Sociology*, pp. 1420–1443, 1978.

[35] M. Granovetter, “Network sampling: Some first steps,” *American Journal of Sociology*, pp. 1287–1303, 1976.

[36] M. Hamdi, V. Krishnamurthy, and G. Yin, “Tracking a Markov-modulated stationary degree distribution of a dynamic random graph,” *IEEE Transactions on Information Theory*, vol. 60, no. 10, pp. 6609–6625, 2014.

[37] J. D. Hamilton, “Time series analysis,” *Princeton University Press*, vol. 2, 1994.

[38] D. D. Heckathorn, “Respondent-driven sampling: A new approach to the study of hidden populations,” *Social Problems*, vol. 44, no. 2, pp. 174–199, 1997.

[39] D. D. Heckathorn, “Respondent-driven sampling II: Deriving valid population estimates from chain-referral samples of hidden populations,” *Social Problems*, vol. 49, no. 1, pp. 11-34, 2002.

[40] M. Hernández-González and M. Basin, “Discrete-time filtering for nonlinear polynomial systems over linear observations,” *International Journal of Systems Science*, vol. 45, no. 7, pp. 1461–1472, 2014.

[41] H. Hethcote, “The mathematics of infectious diseases,” *SIAM Review*, vol. 42, no. 4, pp. 599–653, 2000.

[42] M. O. Jackson, “Social and Economic networks,” *Princeton University Press*, vol. 3, 2008.

[43] M. O. Jackson and B. W. Rogers, “Relating network structure to diffusion properties through stochastic dominance,” *The BE Journal of Theoretical Economics*, vol. 7, no. 1, 2007.

[44] M. O. Jackson and L. Yariv, “Diffusion of behavior and equilibrium properties in network games,” *The American Economic Review*, vol. 97, no. 2, pp. 92–98, 2007.

[45] A. Java, X. Song, T. Finin, and B. Tseng, “Why we Twitter: understanding microblogging usage and communities,” *Proceedings of the 9th WebKDD and 1st SNA-KDD workshop on Web Mining and Social Network Analysis*, pp. 56–65, 2007.

[46] F. Jin, E. Dougherty, P. Saraf, Y. Cao, and N. Ramakrishnan, “Epidemiological modeling of news and rumors on Twitter,” *Proceedings of the 7th Workshop on Social Network Mining and Analysis*, pp. 8, 2013.

[47] V. Krishnamurthy, “Partially Observed Markov Decision Processes,” *Cambridge University Press*, 2016.

[48] V. Krishnamurthy, S. Bhatt, and T. Pedersen, ”Tracking Infection Diffusion in Social Networks: Filtering Algorithms and Threshold Bounds”, *IEEE Transactionson Signal and Information Processing over Networks*, vol. 3, no. 2, 2017.

[49] V. Krishnamurthy, O. N. Gharehshiran, and M. Hamdi, "Interactive Sensing and Decision Making in Social Networks," *Foundations and Trends in Signal Processing*, vol. 7, no. 1, pp. 1–196, 2014.

[50] H. Kushner, "Weak convergence methods and singularly perturbed stochastic control and filtering problems," *Springer Science & Business Media*, 2012.

[51] H. Kwak, C. Lee, H. Park, and S. Moon, "What is Twitter, a social network or a news media?," *Proceedings of the 19th international conference on World wide web*, pp. 591–600, 2010.

[52] S. Lee, "Understanding respondent driven sampling from a total survey error perspective," *Survey Practice*, vol. 2, no. 6, 2013.

[53] J. Leskovec, D. Chakrabarti, J. Kleinberg, and C. Faloutsos, "Realistic, mathematically tractable graph generation and evolution, using Kronecker multiplication," *European Conference on Principles of Data Mining and Knowledge Discovery*, pp. 133–145, 2005.

[54] J. Leskovec and C. Faloutsos, "Sampling from large graphs," *Proceedings of the 12th ACM SIGKDD international conference on Knowledge Discovery and Data Mining*, pp. 631–636, 2006.

[55] J. Leskovec, M. McGlohon, C. Faloutsos, N. Glance, and M. Hurst, "Patterns of cascading behavior in large blog graphs," *In Proceedings of the 2007 SIAM international conference on data mining*, pp. 551–556, 2007.

[56] E. Lieberman, C. Hauert, and M. A. Nowak, "Evolutionary Dynamics on Graphs," *nature*, vol. 433, no. 7023, pp. 312–316, 2005.

[57] D. López-Pintado, "An Overview of Diffusion in Complex Networks," *Springer International Publishing*, 2016.

[58] D. López-Pintado, “Contagion and coordination in random networks,” *International Journal of Game Theory*, vol. 34, no. 3, pp. 371–381, 2006.

[59] D. López-Pintado, “Diffusion in complex social networks,” *Games and Economic Behavior*, vol. 62, no. 2, pp. 573–590, 2008.

[60] G. Lotan,E. Graeff, M. Ananny, D. Gaffney, and I. Pearce, “The Arab Spring the revolutions were tweeted: Information flows during the 2011 Tunisian and Egyptian revolutions,” *International Journal of Communication*, vol. 5, no. 31, 2011.

[61] K. V. Mardia and C. G. Khatri, “Uniform distribution on a Stiefel manifold,” *Journal of Multivariate Analysis*, vol. 7, no. 3, pp. 468–473 1977.

[62] P. L. Odell and A. H. Feiveson, “A numerical procedure to generate a sample covariance matrix,” *Pearson Education India*, vol. 61, no. 313, pp. 199–203, 1966.

[63] M. McPherson, L. Smith-Lovin, and J. M. Cook, “Birds of a feather: Homophily in social networks,” *Annual Review of Sociology*, pp. 415–444, 2001.

[64] A. Mata, R. Ferreira, and S. Ferreira, “Heterogeneous pair-approximation for the contact process on complex networks,” *New Journal of Physics*, vol. 16, no. 5, 2014.

[65] E. Mossel and S. Roch, “On the submodularity of influence in social networks,” *Proceedings of the thirty-ninth annual ACM symposium on Theory of Computing*, pp. 128–134, 2007.

[66] A. Müller and D. Stoyan, “Comparison Methods for Stochastic Models and Risks,” *Wiley*, vol. 389, 2002.

[67] M. A. Nowak, C. Tarnita, and T. Antal “Evolutionary dynamics in structured populations,” *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, vol. 365, no. 1537, pp. 19-30, 2010.

[68] H. Ohtsuki, C. Hauert, E. Lieberman, and M. A. Nowak, “A simple rule for the evolution of cooperation on graphs and social networks,” *Nature*, vol. 441, no. 7092, pp. 502-505, 2006.

[69] R. Olfati-Saber, “Distributed Kalman filtering for sensor networks,” *46th IEEE Conference on Decision and Control*, pp. 5492–5498, 2007.

[70] Z. Papacharissi and M. de F. Oliveira, “Affective news and networked publics: The rhythms of news storytelling on #Egypt,” *Journal of Communication*, vol. 62, no. 2, pp. 266–282, 2012.

[71] R. Pastor-Satorras and A. Vespignani, “Epidemic spreading in scale-free networks,” *Physical Review Letters*, vol. 86, no. 14, pp. 3200, 2001.

[72] L. Perez and S. Dragicevic, “An agent-based approach for modeling dynamics of contagious disease spread,” *International journal of health geographics*, vol. 8, no. 1, pp.50, 2009.

[73] J. Poncela, J. Gómez-Gardenes, J. Moreno, and Y. M. Floría, “Consolidating birth-death and death-birth processes in structured populations,” *International Journal of Bifurcation and Chaos*, vol. 20, no. 3, pp. 849-857, 2010.

[74] M. A. Porter, J. P. Gleeson, “Dynamical systems on networks: A tutorial,” *Springer*, 2016.

[75] L. Shi, “Kalman filtering over graphs: Theory and applications,” *IEEE Transactions on Automatic Control*, vol. 54, no. 9, pp. 2230–2234, 2009.

[76] H. A. Simon, “On a class of skew distribution functions,” *Biometrika*, vol. 42, no. 3, pp. 425–440, 1955.

[77] G. Szabó and G. Fath, “Evolutionary games on graphs,” *Physics reports*, vol. 446, no. 4, pp. 97–216, 2007.

[78] M. Taylor, “Exact and approximate epidemic models on networks: theory and applications,” *University of Sussex*, 2013.

[79] P. Tichavsky, C. Muravchik, and A. Nehorai, “Posterior Cramér-Rao bounds for discrete-time nonlinear filtering,” *IEEE Transactions on Signal Processing*, vol. 46, no. 5, pp. 1386–1396, 1998.

[80] D. Topkis, “Supermodularity and Complementarity,” *Princeton University Press*, 1998.

[81] H. L. Van Trees, “Detection, estimation, and modulation theory,” *John Wiley & Sons*, 2004.

[82] F. Vega-Redondo, “Complex social networks,” *Cambridge University Press*, no. 44, 2007.

[83] Z. Wang, W. Zhang, and C. W. Tan, “On inferring rumor source for SIS model under multiple observations,” *International Conference on Digital Signal Processing (DSP)*, pp. 755–759, 2015.

[84] L. Weng, F. Menczer, and Y. Y. Ahn, “Virality prediction and community structure in social networks,” *Scientific Reports*, vol. 3, 2013.

[85] R. J. Wilson, “An introduction to graph theory,” *Pearson Education India*, 1970.

[86] H. Yang, W. Zhi-Xi, and D. Wen-Bo, “Evolutionary games on scale-free networks with tunable degree distribution,” *EPL (Europhysics Letters)*, vol. 99, no. 1, 2012.

[87] D. H. Zanette and S. Risau-Gusmán, “Infection spreading in a population with evolving contacts,” *Journal of Biological Physics*, vol. 34, no. 1, pp. 135–148, 2008.

[88] J. Zhang and M. F. J. Moura, “Diffusion in social networks as SIS epidemics: beyond full mixing and complete graphs,” *IEEE Journal of Selected Topics in Signal Processing*, vol. 8, no. 4, pp. 537–551, 2014.

Appendix A

Related Proofs

In this Appendix, we provide a proof of the Mean Field Approximation used in Chapter 2.

A.1 Proof of Theorem 1

(Mean Field Dynamics) [48]

Part 1 of the theorem is a standard martingale representation of a Markov chain, see for example [47, page 20].

The proof of the mean field dynamics approximation in [12] is not readily accessible to an engineering reader. We show below that the proof is a simple consequence of Azuma-Hoeffding inequality and Gronwall's inequality.

Theorem 2 (Azuma-Hoeffding Inequality [14])**.** *Suppose $S_T = \sum_{\tau=1}^{T} \zeta_\tau + S_0$ where $\{\zeta_\tau\}$ is a martingale difference process with bounded differences satisfying $|\zeta_\tau| \leq \Delta_\tau$ almost surely where Δ_τ are finite constants. Then for any $\epsilon > 0$,*

$$\mathbb{P}(|S_T - S_0| \geq \epsilon) \leq 2 \exp\left(- \frac{\epsilon^2}{2 \sum_{\tau=1}^{T} \Delta_\tau^2} \right) \quad \square$$

A bound on the deviation between the mean field dynamics $\mathbf{x}_n$ in (2.15) and infected population state $\tilde{\mathbf{x}}_n$ in (2.14) is evaluated in the form of two lemmas, namely, Lemma 1 and Lemma 2 below. Recall that ζ_τ is a $2D$-dimensional finite-state martingale increment process defined in (2.14) with $\|\zeta_\tau\|_2 \leq \frac{\Gamma}{M}$ for some positive constant Γ.

Lemma 1. *Let* $\varphi_n = \boldsymbol{x}_n - \widetilde{\boldsymbol{x}}_n$. *Then* $\|\varphi_n\|$ *satisfies*

$$\|\varphi_{n+1}\|_\infty \le \|\varphi_0\|_\infty + \frac{\beta}{M}\sum_{\tau=1}^{n} \|\varphi_\tau\|_\infty + S_T.$$

Lemma 2. *Let* $S_T = \max_{1\le n\le T} \| \sum_{\tau=1}^{n} \zeta_\tau\|_\infty$. *Then*

$$\mathbb{P}\big(S_T \ge \epsilon\big) \le 2\exp\left(-\frac{\epsilon^2 M^2}{2\Gamma T}\right)$$

Proof of Theorem 1: With Lemmas 1 and 2, the proof of Theorem 1 is as follows. Applying Gronwall's inequality[23] to Lemma 1 yields $\|\varphi_n\|_\infty \le S_T \exp\left[\frac{\beta n}{M}\right]$, which in turn implies

$$\max_{1\le n\le T} \|\varphi_n\|_\infty \le S_T \exp\left[\frac{\beta T}{M}\right].$$

As a result

$$\begin{aligned}\mathbb{P}(\max_{1\le n\le T} \|\varphi_n\|_\infty > \epsilon) &\le \mathbb{P}(S_T \exp\left[\frac{\beta T}{M}\right] > \epsilon) \\ &= \mathbb{P}\big(S_T > \exp\left[-\frac{\beta T}{M}\right] \epsilon\big)\end{aligned}$$

Next applying Lemma 2 to the right hand side yields

$$\mathbb{P}(\max_{1\le n\le T} \|\varphi_n\|_\infty > \epsilon) \le 2\exp\left(-\exp\big(\frac{-2\beta T}{M}\big)\, \epsilon^2 \frac{M^2}{2\Gamma T}\right)$$

Finally choosing $T = c_1 M$, for some positive constant c_1 yields

$$\mathbb{P}(\max_{1\le n\le T} \|\varphi_n\|_\infty > \epsilon) \le 2\exp(-C_2\epsilon^2 M)$$

[23] Gronwall's inequality: if $\{x_n\}$ and $\{b_n\}$ are non-negative sequences and $a \ge 0$, then

$$\bar{x}_n \le a + \sum_{j=1}^{n-1} x_j b_j \implies x_n \le a\exp(\sum_{j=1}^{n-1} b_j)$$

where $C_2 = \exp(-2\beta c_1)\frac{1}{2\Gamma c_1}$. □

Proof of Lemma 1: Define the following $2D-$ dimensional vectors:

$$\mathcal{H}(\mathbf{x}_n) = \mathcal{P}_{21}(\mathbf{x}_n) - \mathcal{P}_{12}(\mathbf{x}_n),$$
$$\mathcal{H}(\widetilde{\mathbf{x}}_n) = \mathcal{P}_{21}(\widetilde{\mathbf{x}}_n) - \mathcal{P}_{12}(\widetilde{\mathbf{x}}_n).$$

Recall from (2.14) and (2.15),

$$\begin{aligned}
\varphi_{n+1} &= \varphi_n + \frac{1}{M}[\mathcal{H}(\mathbf{x}_n) - \mathcal{H}(\widetilde{\mathbf{x}}_n)] + \zeta_n \\
&= \varphi_0 + \frac{1}{M}\sum_{\tau=1}^{n}[\mathcal{H}(\mathbf{x}_\tau) - \mathcal{H}(\widetilde{\mathbf{x}}_\tau)] + \sum_{\tau=1}^{n}\zeta_\tau \\
\|\varphi_{n+1}\|_\infty &\leq \|\varphi_0\|_\infty + \frac{1}{M}\sum_{\tau=1}^{n}\|\mathcal{H}(\mathbf{x}_\tau) - \mathcal{H}(\widetilde{\mathbf{x}}_\tau)\|_\infty + \|\sum_{\tau=1}^{n}\zeta_\tau\|_\infty \\
&\leq \|\varphi_0\|_\infty + \frac{\beta}{M}\sum_{\tau=1}^{n}\|\varphi_\tau\|_\infty + S_T
\end{aligned}$$

The last inequality is justified as follows: From (2.11) and (2.12), $P_{21}(d, \bar{x}_n(d)) = \rho(d)(1 - \bar{x}_n(d))\bar{P}_{21}(d, a)$ and $P_{12}(d, \bar{x}_n(d)) = \rho(d)\bar{x}_n(d)\bar{P}_{12}(d, a)$ for some a, where a is the number of infected neighbors. Hence

$$\mathcal{H}(\mathbf{x}_n, i) - \mathcal{H}(\widetilde{\mathbf{x}}_n, i) \leq \beta(\bar{x}_n(i) - x_n(i)),$$

where $\beta = \max_d \left[\rho(d)(\bar{P}_{21}(d, a) + \bar{P}_{12}(d, a))\right]$ is bounded.

Proof of Lemma 2: $\|\sum_{\tau=1}^{n}\zeta_\tau\|_\infty = \max_i |\sum_{\tau=1}^{n} e_i'\zeta_\tau| = |\sum_{\tau=1}^{n} e_{i^*}'\zeta_\tau|$ for some i^*, where e_i denotes the vector having 1 at the i^{th} position and zero elsewhere. Since $e_{i^*}'\zeta_\tau$ is a martingale difference process with $|e_{i^*}'\zeta_\tau| \leq \sqrt{\Gamma}/M$ applying the Azuma-Hoeffding inequality (Theorem 2) yields

$$\mathbb{P}(\|\sum_{\tau=1}^{n}\zeta_\tau\|_\infty \geq \epsilon) = \mathbb{P}(|\sum_{\tau=1}^{n} e_{i^*}'\zeta_\tau| \geq \epsilon) \leq 2\ \exp\left[-\frac{\epsilon^2 M^2}{2\Gamma n}\right]$$

The right hand side is increasing with n. So clearly,

$$\mathbb{P}(\max_{1\leq n\leq T}\|\sum_{\tau=1}^{n}\zeta_\tau\|_\infty \geq \epsilon) \leq 2\ \exp\left[-\frac{\epsilon^2 M^2}{2\Gamma T}\right]$$

□