Modeling and analysis of epidemic processes over large networks from limited data

by

Sifat Afroj Moon

B.S., Bangladesh University of Engineering and Technology, 2013

M.S., Kansas State University, 2018

AN ABSTRACT OF A DISSERTATION

submitted in partial fulfillment of the
requirements for the degree

DOCTOR OF PHILOSOPHY

Mike Wiegers Department of Electrical and Computer Engineering
Carl R. Ice College of Engineering

KANSAS STATE UNIVERSITY
Manhattan, Kansas

2021

Abstract

Networks are ubiquitous in today's interlinked world, allowing various types of flow along with their links, for instance, rumor, knowledge, norms (social networks), electricity (power grid), money (financial networks), goods (trade networks), and infectious pathogens (disease networks). In particular, network epidemic models offer analytical and numerical platforms to quantify and forecast the transmission of a pathogen. In this context, a network is a structure of nodes and links, where nodes can represent individuals, locations, or groups, and links can represent physical contacts between individuals, movement flows between locations, or interaction between groups. Studying infectious disease dispersal with network epidemic models is a powerful application of network science.

An extremely challenging aspect of studying infectious pathogen dispersal is the insufficient knowledge of the underlying network or unknown epidemic model parameters. Sometimes networks are unattainable because of data privacy, lack of data integrity, missing data, or lack of higher resource requirements. This dissertation provides a guideline to study epidemic processes on a network from limited available data.

In the USA, no mandatory livestock regulatory system exists because of a cultural preference for privacy. This dissertation proposes a general algorithm to develop a livestock movement network from the limited available data to fill this gap. In this network, nodes represent livestock subpopulations, and links represent livestock directional movements between subpopulations. Network centrality measures are beneficial to understand the contact pattern of a movement network and can assist in detecting the superspreader nodes, which play a critical role in the movement flows and disease transmission. Understanding the role of superspreaders in a movement network is useful for policymakers to control disease outbreaks efficiently. This is possible because the network centrality analysis in the livestock movement network reveals small-world phenomena in the US livestock industry.

Individual-based network models are becoming popular due to their capability to integrate heterogeneous social mixing. However, individual contact networks are not available because of privacy concerns. This dissertation offers an age-specific multilayer individual-based contact network developed from demographic data and Google mobility data. Combining this network with an epidemic model led to costs and benefits of contact tracing being investigated as a key mitigation strategy in the COVID-19 transmission. Then, an approximate Bayesian computation based on a sequential Monte Carlo sampling (ABC-SMC) method allowed network models to estimate the disease propagation rate from the COVID-19 incidence data. The ABC-SMC method is ideal for parameter estimation and model selection of a complex system when the likelihood function is intractable or computationally expensive to evaluate. This work provides a general, flexible, and complete framework to study an epidemic process from data at the individual level.

Some individual contact networks have an enormous set of nodes/agents; however, individual-based stochastic epidemic modeling, like the generalized epidemic modeling framework (GEMF), over those vast networks is computationally expensive and time-consuming. This dissertation proposes a group-based continuous-time Markov epidemic model framework to reduce the computational time of the individual-based framework (GEMF) by reducing the state-space in the Markov chain. The number of states in the individual-based Markov model is M^N (where M is the number of compartments, and N is the number of nodes), and it increases exponentially with the number of nodes N. By partitioning the nodes into C disjointed groups, the group-based approach reduces the state-space to $\left[\prod_{i=1}^{C} \binom{N_i+M-1}{M-1} \right]$, which is already polynomial in N for a constant number of groups and quasi-polynomial in N for a logarithmic number of groups; i.e., $C = O(\log N)$. Here, N_i represents the number of nodes in a group i and $i = 1, 2, ..., C$. The simulation results reveal that the accuracy of the group-based approach depends on the network structures and grouping approaches.

In summary, this dissertation enhances the current knowledge of network epidemic models both in application and theory; therefore, it can serve as a foundation work of follow-on efforts related to the network epidemic modeling over large networks.

Modeling and analysis of epidemic processes over large networks from limited data

by

Sifat Afroj Moon

B.S., Bangladesh University of Engineering and Technology, 2013

M.S., Kansas State University, 2018

A DISSERTATION

submitted in partial fulfillment of the
requirements for the degree

DOCTOR OF PHILOSOPHY

Mike Wiegers Department of Electrical and Computer Engineering
Carl R. Ice College of Engineering

KANSAS STATE UNIVERSITY
Manhattan, Kansas

2021

Approved by:

Major Professor
Caterina M. Scoglio

Copyright

Abstract

Networks are ubiquitous in today's interlinked world, allowing various types of flow along with their links, for instance, rumor, knowledge, norms (social networks), electricity (power grid), money (financial networks), goods (trade networks), and infectious pathogens (disease networks). In particular, network epidemic models offer analytical and numerical platforms to quantify and forecast the transmission of a pathogen. In this context, a network is a structure of nodes and links, where nodes can represent individuals, locations, or groups, and links can represent physical contacts between individuals, movement flows between locations, or interaction between groups. Studying infectious disease dispersal with network epidemic models is a powerful application of network science.

An extremely challenging aspect of studying infectious pathogen dispersal is the insufficient knowledge of the underlying network or unknown epidemic model parameters. Sometimes networks are unattainable because of data privacy, lack of data integrity, missing data, or lack of higher resource requirements. This dissertation provides a guideline to study epidemic processes on a network from limited available data.

In the USA, no mandatory livestock regulatory system exists because of a cultural preference for privacy. This dissertation proposes a general algorithm to develop a livestock movement network from the limited available data to fill this gap. In this network, nodes represent livestock subpopulations, and links represent livestock directional movements between subpopulations. Network centrality measures are beneficial to understand the contact pattern of a movement network and can assist in detecting the superspreader nodes, which play a critical role in the movement flows and disease transmission. Understanding the role of superspreaders in a movement network is useful for policymakers to control disease outbreaks efficiently. This is possible because the network centrality analysis in the livestock movement network reveals small-world phenomena in the US livestock industry.

Individual-based network models are becoming popular due to their capability to integrate heterogeneous social mixing. However, individual contact networks are not available because of privacy concerns. This dissertation offers an age-specific multilayer individual-based contact network developed from demographic data and Google mobility data. Combining this network with an epidemic model led to costs and benefits of contact tracing being investigated as a key mitigation strategy in the COVID-19 transmission. Then, an approximate Bayesian computation based on a sequential Monte Carlo sampling (ABC-SMC) method allowed network models to estimate the disease propagation rate from the COVID-19 incidence data. The ABC-SMC method is ideal for parameter estimation and model selection of a complex system when the likelihood function is intractable or computationally expensive to evaluate. This work provides a general, flexible, and complete framework to study an epidemic process from data at the individual level.

Some individual contact networks have an enormous set of nodes/agents; however, individual-based stochastic epidemic modeling, like the generalized epidemic modeling framework (GEMF), over those vast networks is computationally expensive and time-consuming. This dissertation proposes a group-based continuous-time Markov epidemic model framework to reduce the computational time of the individual-based framework (GEMF) by reducing the state-space in the Markov chain. The number of states in the individual-based Markov model is M^N (where M is the number of compartments, and N is the number of nodes), and it increases exponentially with the number of nodes N. By partitioning the nodes into C disjointed groups, the group-based approach reduces the state-space to $\left[\prod_{i=1}^{C} \binom{N_i+M-1}{M-1} \right]$, which is already polynomial in N for a constant number of groups and quasi-polynomial in N for a logarithmic number of groups; i.e., $C = O(\log N)$. Here, N_i represents the number of nodes in a group i and $i = 1, 2, ..., C$. The simulation results reveal that the accuracy of the group-based approach depends on the network structures and grouping approaches.

In summary, this dissertation enhances the current knowledge of network epidemic models both in application and theory; therefore, it can serve as a foundation work of follow-on efforts related to the network epidemic modeling over large networks.

Table of Contents

List of Figures

List of Tables

Acknowledgments

The work presented in this dissertation would not have been possible without the support of the wonderful people around me. I take this opportunity to express my sincere gratitude and appreciation to all those who made this dissertation possible.

First and foremost, I would like to thank my advisor, Dr. Caterina M. Scoglio, for her continuous support, guidance, and encouragement. She always made time to listen to any of my queries and tried her best to guide me compassionately. Her passion and enthusiasm kept me continuously engaged with my research.

I would like to express my appreciation to my co-advisor, Dr. Lee W. Cohnstaedt, for his valuable and fascinating advice throughout my Ph.D. journey. He always encouraged me to think out of the box, and his personality inspired me to be a life-long learner.

My sincere gratitude to my advisory committee members Dr. Don Gruenbacher, Dr. Punit Prakash, and Dr. William Hsu, for their time and constructive comments to improve this work. I would like to express my sincere appreciation to the Electrical and Computer Engineering department for all of its supports.

It has been my privilege to collaborate with Dr. Faryad Darabi Sahneh during my Ph.D. I have learned many great things from him. I also wish to thank Dr. Christopher Mundt, PI of the EEID project, for all of his kind supports.

Special thanks are extended to my officemates at the NetSE research lab. Their collaboration and friendship shall always be remembered. I also wish to thank my friends at K-State for sharing happy moments and being with me.

I would like to appreciate my amazing family members. I am blessed to have the most caring and supportive family in the world — special thanks to my parents, Abdul Gofur and Aleya Khatun, for their never-ending love and kindness. I would also like to take the name of my siblings, Jahangir Alam, Mahbubul Alam, and Ellora Yasi, who always trust me and try their best to encourage me.

Especial acknowledgments to my amazing husband, Tanvir Ferdousi, for his unlimited support and care, and my lovely daughter, Ava, for her innocent love, which can make me forget anything. Words are not capable of expressing my gratitude for always being there for me through thick and thin.

Dedication

To my mother, Aleya Khatun, and my daughter, AVA.

Preface

This dissertation is submitted for partial fulfillment of the requirements for the degree Doctor of Philosophy in the Department of Electrical and Computer Engineering at Kansas State University. The thesis work was conducted from August 2016 to May 2021 under the supervision of Professor Caterina Scoglio.

This work is, to the best of my knowledge, original, except where acknowledgments and references are made to previous work. Part of this work has been presented in the following publications.

Peer-reviewed journal articles

1. Sifat A Moon and Caterina M Scoglio. Contact tracing evaluation for COVID-19 transmission in the different movement levels of a rural college town in the USA. *Scientific Reports*, 11(1):1–12, 2021. (impact factor: 3.998)

2. Sifat A Moon, Faryad Darabi Sahneh, and Caterina M Scoglio. Group-based general epidemic modeling for spreading processes on networks: GroupGEM. *IEEE Transactions on Network Science and Engineering*, 8(1):434–446, 2021. doi: 10.1109/TNSE. 2020.3039494. (impact factor: 5.213)

3. Sifat A Moon, Tanvir Ferdousi, Adrian Self, and Caterina M Scoglio. Estimation of swine movement network at farm level in the US from the census of agriculture data. *Scientific Reports*, 9(1):6237, 2019. (impact factor: 3.998)

4. Sifat A Moon, Lee W Cohnstaedt, D Scott McVey, and Caterina M Scoglio. A spatiotemporal individual-based network framework for West Nile virus in the USA: spreading pattern of West Nile virus. *PLOS Computational Biology*, 15(3):e1006875, 2019. (impact factor: 4.801)

5. Tanvir Ferdousi, Sifat Afroj Moon, Adrian Self, and Caterina Scoglio. Generation of swine movement network and analysis of efficient mitigation strategies for African swine fever virus. *PLOS ONE*, 14(12), 2019. (impact factor: 2.74)

The conducted research has been supported by the NSF\NIH\USDA\BBSRC Ecology and Evolution of Infectious Diseases (EEID) Program through USDA-NIFA Award 2015-67013-23818 and by the State of Kansas, National Bio and Agro-Defense Facility (NBAF) Transition Fund through the National Agricultural Biosecurity Center (NABC) at Kansas State University.

Chapter 1

Introduction

1.1 Background

Networks have the ability to represent the structure and function of various physical, social, environmental, and technological systems. A network-based epidemic model can describe a dynamic process such as information dissemination, cultural norms, or viruses throughout a network [6, 7]. Understanding the epidemic processes on networks allows us to assess risks, plan for effective control measures, and identify superspreaders. Modeling and analysis of an epidemic process on a network consist of four basic steps:

1. Inference of the network for interacting agents of a system from the data
2. Definition of a compartmental epidemic model to describe the state of a node from the disease dynamics
3. Estimation of the disease model parameters from the incidence data or experiments
4. Decisions concerning a modeling approach

1.1.1 Network estimation

Regarding real-world epidemic phenomena, the exact network structure is unknown to the researchers; therefore, inferring the network structure from the limited data is crucial. A network of

a system depends on the purpose of the research; for example, in a livestock movement network, nodes can be individual farms, whereas links or edges represent livestock movement between two farms. Then again, nodes can be individual animals, and edges represent direct contact between two animals. Estimation of a network has many challenges: for example, lack of data integrity, maintenance of data anonymity, and identification of missing links or data. In the literature, researchers often infer the network for an epidemic modeling from the disease incidence data or knowledge of the underlying system [8–12].

1.1.2 Compartmental epidemic model

To represent the incidence, transmission, and persistence of infectious disease, defining a proper realistic epidemic model is important. Therefore, compartmental epidemic models have been widely used in the study of epidemic processes [7, 13]. In a compartmental epidemic model, the population distributes into different compartments; for example, members can be susceptible, infected, or recovered.

Moreover, compartmental models can incorporate various disease dynamics. Some dynamics follow the susceptible-infected-susceptible (SIS) [14, 15] or susceptible-infected-recovered (SIR) [13, 16, 17] compartmental model, while some need more complex compartmental models. In the SIS model, a susceptible node transition to infected based on the infection rate, and an infected node becomes susceptible with a given recovery rate. Therefore, in the SIS model, a node can become infected and susceptible several times because there is no immunity, whereas, in the SIR model, a susceptible node may become infected then recover and be removed from becoming susceptible again. SIS and SIR models are the basis of most other more complicated disease models [6].

A model should have an appropriate balance between accuracy and complexity for optimal usefulness [7]. Accuracy is the ability to reproduce the observed data, adding all the relevant factors of the disease dynamics. However, too many factors or compartments increase model complexity, which makes model optimization difficult. In addition, too many factors can make the model vulnerable to overfitting [11, 18].

1.1.3 Parameter estimation

In a compartmental epidemic model, a node can move from one compartment to another compartment via a transition rate, which is defined as a parameter of an epidemic model [7, 13]. The transition rate can be fixed, or temporal [6]. Estimation of a transition rate is possible from incidence data or laboratory experiments. In the network epidemic modeling, transition rates are characterized by two types: nodal transition and edge-based transition [1, 19]. A nodal transition of a node depends on the current state or compartment of the node. An edge transition of a node depends on the current state of the node and the state of its neighboring nodes. Each edge transition has an influencer compartment that is the compartment of the neighboring nodes or state of the neighboring nodes, which affects the edge transition. For example, in a susceptible-infected-susceptible epidemic model, the susceptible-to-infected edge transition of a susceptible node is caused by its infected neighboring nodes. Therefore, the infected compartment is the influencer compartment for this edge transition. Both nodal transition and edge-based transition rates are possible to estimate by fitting a mathematical model to incidence or prevalence data [20–22].

1.1.4 Network epidemic modeling approaches

Researchers have developed several tools to model epidemic processes on networks depending on the research questions and resources; some are very detail-oriented, and some are not [6, 14, 23–27]. In this section, we will discuss three relevant epidemic modeling approaches.

In the individual-based approach [1, 14, 16, 23], nodes are at the individual level where links represent the connection between two individuals. The individual-based approach preserves the full description of a network, which contains detailed information about the exact neighborhood.

The metapopulation method [25–27] is another approach to model epidemic processes. In this model, nodes are not individual entities; rather, a node is an entity or a place where multiple individuals can be located. Therefore, a node represents a subpopulation. The movement of individuals establishes links between two nodes or subpopulations. The metapopulation model considers that individuals mix homogeneously in a node and diffuse from one node to another with

a fixed mobility rate. For this reason, the metapopulation model loses all the information inside a subpopulation.

The group-based approach [19, 28] provides a balance between the individual-based approach and the metapopulation approach. In the group-based approach, nodes are divided into several disjoint partitions, where a group represents a subnetwork. The group-based approach preserves certain network properties inside a group.

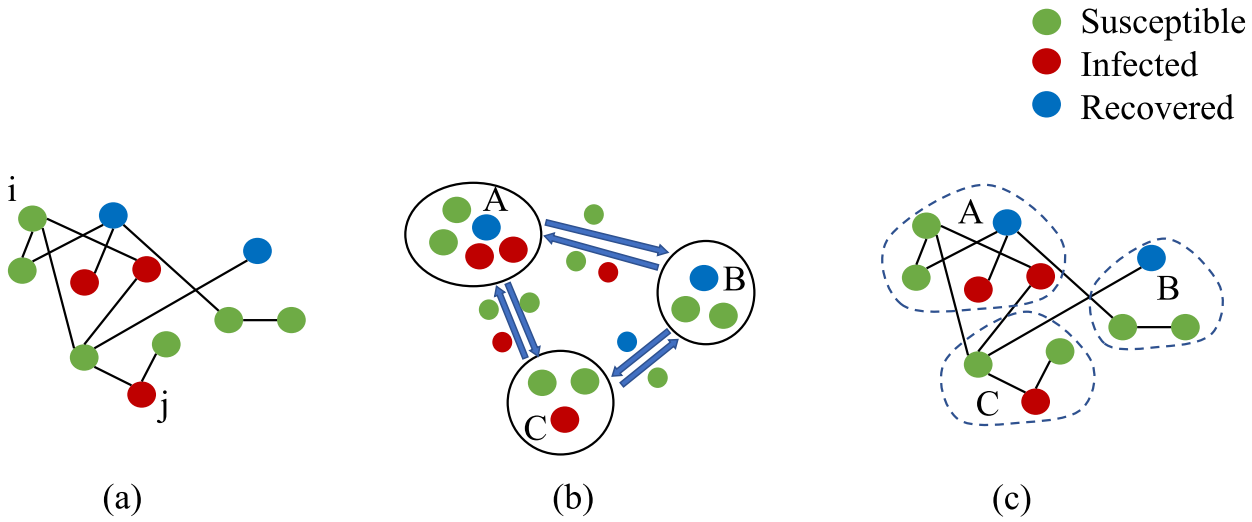


Figure 1.1: **A summary of three different epidemic modeling approaches to model susceptible-infected-recovered (SIR) epidemic process on a network;** (a) Individual-based approach: each node represents each individual, (b) Metapopulation approach: each node represents a collection of individuals, and (c) Group-based approach: nodes are divided into three disjoint partitions, each group represents a subnetwork.

Stochasticity is an essential feature of real-world epidemic phenomena. If it is possible to re-run a real-world epidemic phenomenon, we will not get an identical set of the infected population. Continuous-time Markov processes can describe stochastic epidemic processes on networks in these three epidemic approaches.

1.2 Research motivation

Modeling and simulation are convenient when the resource requirements for data collection and experimental studies are prohibitively high or unattainable. Since epidemic modeling is usually used to understand and predict an infectious disease transmission, understanding the epidemic

process can help policymakers plan effective disease control measures while forecasts can guide experts during tough policy decisions [7].

A network is a useful structure in the study of any spreading phenomenon. Moreover, estimating movement networks from limited data is an important step to model an epidemic process. For example, in the livestock industry, animal movement is one of the major causes of disease dispersal among farms. Thus, a livestock movement network can increase the feasibility of planning effective mitigation strategies that can reduce the risk of disease dispersal. Although in Europe several well-established animal tracking systems are in play, the US has no comprehensive mandatory livestock tracking system because of a cultural preference for privacy and competition among producers [8]. To support the researchers modeling livestock diseases, we propose a framework to develop a farm-level directed weighted movement network (V, E, W) from the publicly available inventory and sales data. The term V denotes the set of farms, E represents the set of links or connections among individual farms, and W denotes the weight of each link. Links or connections among farms represent livestock movement. The weight of a link represents the volume of movement occurring from one farm to another. In this research, we extend the work of Schumm et al. [29], where Schumm et al. [29] have developed a convex optimization problem to identify the cattle county-level movement probabilities from the USDA-NASS data by using the maximum entropy approach. We propose a novel algorithm to develop a farm-level movement network from the county-level movement probabilities. We also adapt the model of Schumm et al. [29] for another livestock industry, swine, which has different population structures and data.

One of the important tasks of epidemic models is to uncover basic epidemiological characteristics of a pathogen by using statistical data analysis. An epidemic model can also provide a means of understanding the effectiveness of a potential control strategy. Accordingly, we have developed an individual-level multilayer heterogeneous network model to find the disease transmission parameters of COVID-19 and to understand the effectiveness of contact tracing in the reopening phase of the USA. The COVID-19 disease is caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) and has created a global health emergency. COVID-19 has affected the lives of billions of people from 2019-2021. To enable prediction of and ultimately policy based on pathogen transmission, individual-based contact-network models are a powerful tool to model

COVID-19 transmission due to its person-to-person transmission nature. The individual-based approach provides the flexibility to represent the heterogeneous social mixing in a community. It also allows us to include a mitigation strategy in the model at the individual level, such as contact tracing. To demonstrate, we have adapted approximate Bayesian computation based on a sequential Monte Carlo Sampling (ABC-SMC) scheme [30–35] for network modeling to estimate the disease parameters from Riley County, Kansas incidence data.

In a network with N nodes, the total number of states in the individual-based Markov chain for a M compartmental epidemic model is M^N [14]. The individual-based continuous-time M^N state Markov model is exact, allowing us to understand the local dynamics of an epidemic. This approach is also more flexible in terms of the initial conditions. However, a drawback of the individual-based approach is longer computational time. To reduce the computational time of the individual-based approach, sometimes researchers scale their population by considering several individuals or a group as a single individual node [11, 36, 37]. This type of scaling can alter the actual system, and estimation of the dynamics can be misleading. This dissertation proposes a group-based general epidemic modeling framework (GroupGEM) [19] to reduce the computational time of the individual-based framework (GEMF) [1] while retaining its advantages. The GroupGEM reduces the state-space of the individual-based Markov process from M^N to $\left[\prod_{i=1}^{C}\binom{\mathcal{N}_i+M-1}{M-1}\right]$ for a partition with C disjoint groups in a network with N nodes, where, $\mathcal{N}_i$ represents the number of nodes in a group i, and $i = 1, 2, ..., C$. Fig 1.2 and 1.3 represent state-space of the individual-based and group-based Markov model for a network with four nodes for an SIS epidemic model.

1.3 Results overview

This dissertation presents a general guideline to model dispersal phenomena on a network from limited available data. The dispersal of rumors, computer viruses, social behavior, and cultural norms can be modeled as an epidemic process. In light of the opportunities, this dissertation will provide guidelines to the researchers of various fields to model stochastic dispersal incidences.

At first, this dissertation develops a general framework to estimate movement in networks from inventory and sales data, flexible enough to apply to other systems, for example, trade networks.

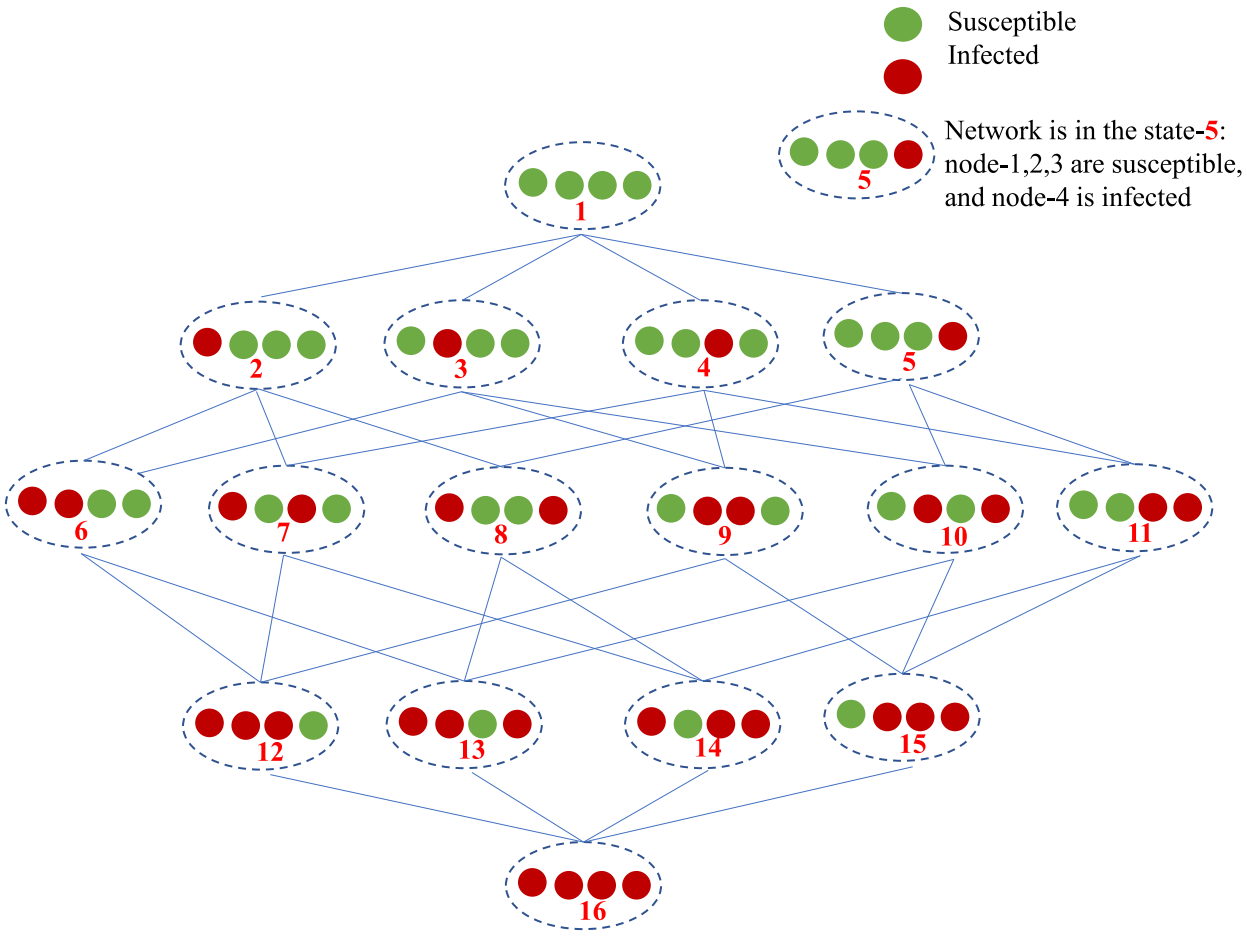


Figure 1.2: **State diagram of an individual-based Markov model for a susceptible-infected-susceptible (SIS) epidemic process.** The network has four nodes; therefore, the individual-based network model has 16 (or 2^4) states.

We also provide remarks on the generated network based on network centrality measures [38]. Future researchers can use the generated network information to study infectious disease transmission in the livestock industry.

This dissertation infers an age-specific individual-based contact network from demographic data and Google mobility data to model COVID-19 transmission. However, this approach is useful to model any dispersal phenomena that is related to direct contacts. By developing a risk assessment framework, we provide a platform to investigate the costs and benefits of a control measure at the individual level. The ABC-SMC method is adapted for network models to estimate unknown parameters and network models from incidence data. ABC-SMC is a computational method of Bayesian statistics that combines a particle filtering method with summary statistics. Thus, this

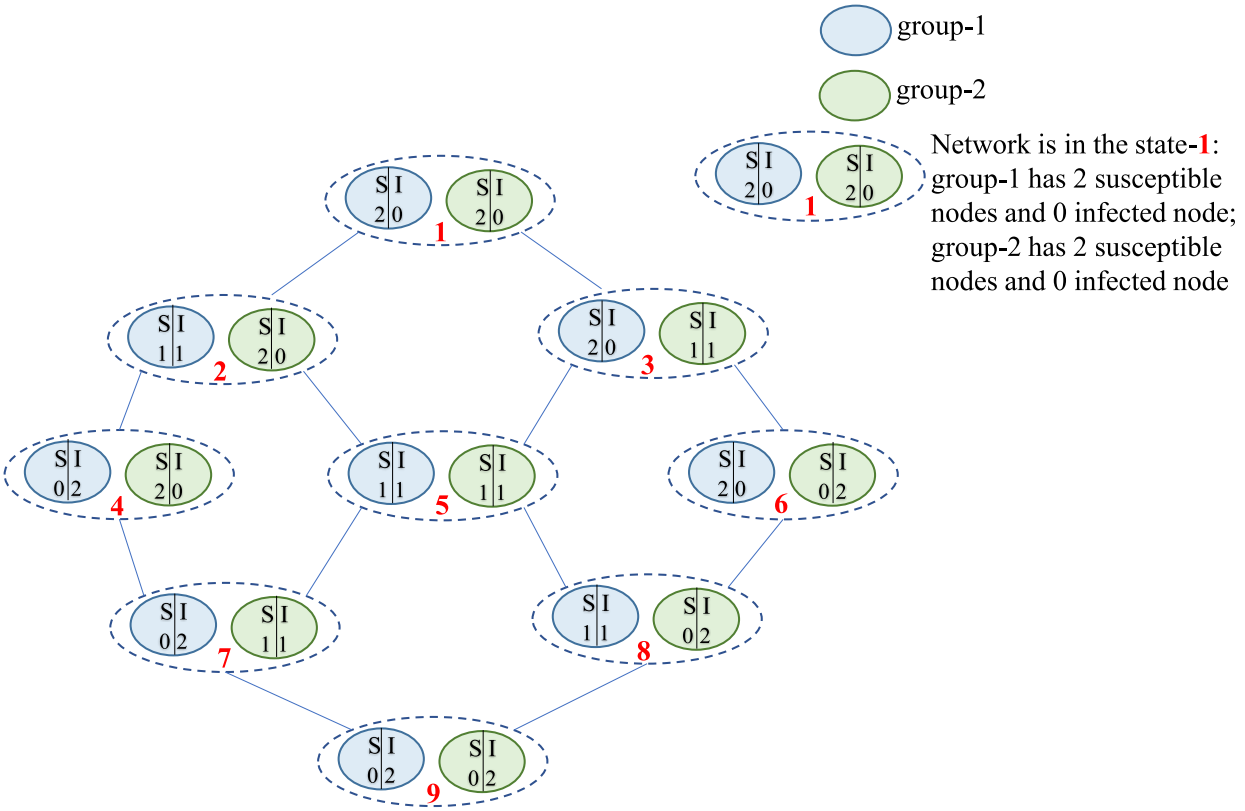


Figure 1.3: **State diagram of a group-based Markov model for a susceptible-infected-susceptible (SIS) epidemic process.** The network has four nodes, which are divided into two disjoint groups; therefore, the group-based network model has 9 states instead of 16.

method is ideal for a stochastic complex model where the likelihood function is intractable or computationally expensive to evaluate.

Finally, we develop a computationally efficient analytical group-based network tool to understand the local dynamics of epidemic processes over very large networks. This work contributes positively to society in understanding how to prevent large-scale catastrophes, including outbreaks of infectious diseases, the propagation of computer viruses, and cascading failure in power-grids. Besides all these opportunities, this research also develops a new package for stochastic numerical simulations.

1.4 Contributions

Below is a summary of the major contributions of this dissertation:

- Develops a convex optimization problem to estimate swine movement probabilities at the county level from comprehensive anonymous inventory and sales data published by the United States Department of Agriculture - National Agriculture Statistics Service database by using the maximum entropy approach.
- Proposes a novel algorithm to develop a directed weighted network at the farm level from the county level movement probabilities.
- Investigates the benefits and costs of contact tracing related to COVID-19 transmission in the reopening process of the USA.
- Infers a general individual-based contact network capable of representing the heterogeneous social mixing and age-structure from demographic data.
- Implements contact tracing in a two-layer network model, which comprises the contact network in the first-layer and the tracing network in the second-layer.
- Adapts approximate Bayesian computation based on sequential Monte Carlo sampling (ABC-SMC) approach for individual-based network model.
- Estimates unknown model parameters from time-series COVID-19 incidence data.
- Proposes a network-based epidemic modeling framework, group-based general epidemic modeling (GroupGEM) framework, to reduce the computational time of individual-based framework (GEMF) while retaining its advantages.
- Develops a continuous-time Markov model for the GroupGEM framework and derives corresponding Kolmogorov equations.
- Implements two different mean-field approximations of the GroupGEM framework to reduce the state-space size further: 1) inter-group mean-field approximation, 2) intra- and inter-group mean-field approximation.
- Extends the GroupGEM framework to multilayer networks.

1.5 Dissertation organization

This dissertation is divided into four significant chapters (in addition to this introduction), which deal with modeling epidemic processes on a network. Each chapter is further subdivided into methods, applications, simulation results, and discussions.

Chapter 2 introduces an approach to estimate a movement network from limited available data by using a maximum entropy approach with application in the swine industry. To enable understanding of the movement network, it uses network centrality measures. The analysis of this network has found evidence of small-world phenomena. Our study suggests that the swine industry in the USA may be vulnerable to infectious disease outbreaks because of the small-world structure of its movement network.

Chapter 3 investigates the benefits and costs of contact tracing in the COVID-19 transmission in the reopening process of a rural college town in the USA. This chapter has developed an individual-based 2-layer contact network model capable of representing the heterogeneous age-specific social mixing. We have used a SEICR (susceptible-exposed-infected-confirmed- removed) epidemic model to emulate COVID-19 transmission. To estimate the unknown parameters from the time-series COVID-19 incidence cases in Riley County, Kansas, we have developed approximate Bayesian computation based on the sequential Monte Carlo Sampling (ABC-SMC) method. Another example of the ABC-SMC method for the West Nile Virus (WNV) is presented in Appendix A. In this research, we investigate the optimum traced percentage for different movement levels. Our simulation finds that the quarantined susceptible people increase with the percentage of traced contacts; however, after a certain number of traced contacts, the number of quarantined susceptible people starts to decrease with the increase in the percentage of traced contacts for any cases [39].

We have developed a general group-based epidemic modeling framework: GroupGEM in Chapter 4, a continuous-time Markov model. This generalization covers any compartmental epidemic models, any static networks (e.g., directed, undirected, weighted), and any disjointed network partitions. The GroupGEM framework has lower computational complexity and faster simulation time than the general individual-based GEMF framework because of the reduced-state

space size. In this chapter, we derive the corresponding Kolmogorov differential equations for the GroupGEM framework. We also propose two mean-field approximation approaches of this framework to reduce the state-space size further. Finally, we have extended the GroupGEM to multilayer networks. We also provide simulation results to investigate the accuracy of the group-based frameworks compared to the individual-based framework in synthetic networks and empirical networks [19].

Closing remarks with future directions on this research are reported in Chapter 5.

Chapter 2

Estimation of swine movement network at farm-level in the USA from the Census of Agriculture data[1]

Swine movement networks among farms/operations are an important source of information to understand and prevent the diseases transmission, nearly nonexistent in the United States. An understanding of the movement networks can help the policymakers in planning effective disease control measures. The objectives of this work are: 1) estimate swine movement probabilities at the county level from comprehensive anonymous inventory and sales data published by the United States Department of Agriculture - National Agriculture Statistics Service database, 2) develop a network based on those estimated probabilities, and 3) analyze that network using network science metrics. First, we use a probabilistic approach based on the maximum information entropy method to estimate the movement probabilities among different swine populations. Then, we create a swine movement network using the estimated probabilities for the counties of the central agricultural district of Iowa. The analysis of this network has found evidence of the small-world phenomenon. Our study suggests that the US swine industry may be vulnerable to infectious disease outbreaks because of the small-world structure of its movement network. Our system is easily

[1]This chapter is a slightly modified version of our published article [12], Copyright ©2019, Scientific Reports.

adaptable to estimate movement networks for other sets of data, farm animal production systems, and geographic regions.

2.1 Background

Livestock are often moved between facilities to reduce costs and improve productivity. There is an old adage, "Livestock follow the grain". Even now this aphorism seems true, as shipping animals is less expensive than shipping grains, which are required for animals to attain their slaughter weights. The corn-belt region (Iowa, Missouri, Illinois, Indiana, and Ohio states) is the largest market for feeder pigs [40] because they are the largest producers of two major sources of hog rations (corn and soybeans). Although movements in the livestock industry can reduce the cost of production, movements have a major role in the risk of pathogens spread. Movement of swine among farms is one of the major pathways for the spread of several diseases (e.g., Porcine reproductive and respiratory syndrome-PRRS, Porcine epidemic diarrhea-PED etc.) in the United States (US) swine industry [41, 42]. Knowledge of livestock movement can be useful in the control of pathogen transmission. In Europe, there are several well-established animal tracking systems. However, similar programs are yet to be mandated for the US. In the US, a comprehensive livestock tracking system has not been implemented because of a cultural preference for privacy and competition between producers [8]. The United State Department of Agriculture (USDA) collects movement information when livestock shipments cross state boundaries. There is no program that collects movement information at the county- or farm-level.

In the prior literature, several models have been developed to understand swine movement in different regions of the US [8–10]. However, all of them used confidential incomplete datasets, which are not publicly accessible, and also which are not inclusive of the whole US. Yadav et al. [10] developed a model to understand classical swine fever outbreak-related outcomes in Indiana. They used data from USAHerds (US Animal Health Emergency Reporting and Diagnostic System), where import-export activities, location of import origin, receiving swine premises, shipment size and shipment date are listed. However, only 22% of the states participates in the USAHerds program. Another research group predicted movement networks of the swine industry for some

counties of Minnesota using a machine learning approach [9]. They used confidential survey data from two counties to train their model. The objective of our research is to understand the swine movement network in the US from publicly available data. A network is a useful structure in the study of any spreading phenomena, where farm-level animal movement networks are used as a key component in the area of pathogen spreading [43, 44].

In this work, we estimate the swine movement probabilities between counties based on published inventory and sales data from the USDA Census of Agriculture. We develop a convex optimization problem with some linear constraints for the US swine industry. To solve this problem, we adapt the cattle movement model from Schumm et al. [29] for the swine population. In particular, we maximize the entropy of the distributions of the objective function (Eq. 2.1). Maximum information entropy methods have been used in various research fields [45–47]. The maximum entropy principle states that the best way to approximate the unknown distribution that satisfies all the constraints will have the maximum entropy [48].

We propose a novel algorithm to develop a farm-level swine movement network using the estimated swine movement probabilities. In this network, nodes (or vertices) represent swine-farms and directed links (or edges or connections) represent directional swine movements between the farms. Network realizations from the interactions among the elements of different dynamic systems can be seen several times in the literature; for example, weighted network for worldwide air transportation[49], network for collaboration among scientists [49], network to understand complex intercellular interactions [50], and network to represent interplay among different physiological systems [51–54]. To understand the generated swine movement network, we use network centrality measures. They have been used often in the literature to understand the livestock movement patterns [55–57]. The network centrality measures can assist in detection of the important farms, which can control the movement flows in the network. This information can be useful to plan effective mitigation strategies to reduce an epidemic size. In the literature, researchers have used targeted vaccination, or quarantine, or culling of important agents to control epidemics [58, 59]. The network centrality measure also can help us to understand the movement pattern. From the analysis of the developed swine movement network, we find a trace of the small world phenomenon and the presence of hubs in the US swine movement network.

First, we develop a convex optimization problem to estimate swine movement probabilities. Next, we propose an algorithm to develop a network based on those probabilities, where nodes or vertices are farms or operations and edges among them represent swine movement. Finally, we analyze the network using different network analysis metrics.

2.2 Data

We have collected the hog inventory, sales, slaughter, and dead/lost pig data from the United States Department of Agriculture National Agricultural Statistics Service (USDA-NASS) [60]. The USDA-NASS conducts a census every five years, which compiles a uniform, comprehensive agricultural data set for each county of the entire US. We used the data from the 2012 Census of Agriculture, as the census of 2017 is not published fully at the time of this research. For each county, two sets of data are available: 1) inventory and 2) sales. In both types, pigs are grouped into seven classes based on operation/farm size. These groups are: size1 (1-24 pigs), size2 (25-49 pigs), size3 (50-99 pigs), size4 (100-199 pigs), size5 (200-499 pigs), size6 (500-999 pigs), and size7 (more than 1000 pigs). For each size group, data for the number of operations and the number of pigs are available. However, several data points are not published to maintain anonymity; we estimate those to develop the network model. The study time of this research is the year 2012. We have assumed that the inventory sizes are constant throughout the year because of the resolution limitation of the available data. Another set of missing data are the geographic farm locations; we use geographical county centroids to measure the distances among counties.

We estimate the swine movement probabilities among sub-populations for the State of Iowa, where a sub-population is denoted as the swine population in a size group in a county. Iowa has the largest swine inventory (31.43%) in the US [61]. In the list of America's top 100 pig farming counties, 42 counties are from Iowa alone[62]. It is also the most vulnerable state for the introduction of classical swine fever and African swine fever viruses due to legal import of live swine [63]. Iowa has 99 counties in total, the number of swine sub-populations in our optimization problem is 99×7.

2.3 Swine movement probability estimation

To estimate the pig movement probabilities in a week among different sub-populations, we use a convex optimization problem. This convex optimization problem consists of two steps: 1) estimation of the non-disclosed data points in the inventory and sales data and 2) estimation of movement probabilities among different sub-populations.

To estimate non-disclosed points in the inventory data, we formulate an entropy function. By maximizing this function, we estimate the data points with minimum assumptions [64]. This process is detailed in Schumm et al[29]. In step 2, we construct a convex optimization problem, which includes a series of linear constraints. The purpose of this problem is to maximize the entropy of the distributions of the objective function, the distributions of the objective function for a sub-population are presented in Fig A.2. The maximum entropy is a well-known method of statistical inference, which has been used in diverse research fields including ecology, thermodynamics, economics, forensics, language processing, astronomy, image processing etc. [47, 65, 66]. This method produces the least biased predictions while maintaining prior knowledge constraints.

In the convex optimization problem, there are C counties and each county has I size groups. A pig from a sub-population can be moved to a sub-population in the state, or moved outside of the state, or not moved at all, or slaughtered, or lost. Therefore, a pig in a sub-population has five movement options, which construct the distributions of the objective function We define the objective function of this estimation problem as,

$$\begin{aligned} max\{Entropy\} = max\{ &- \sum_{x\in C}\sum_{i\in I}\sum_{y\in C}\sum_{j\in I} m^d_{i,j,dist(x,y)} * log(m^d_{i,j,dist(x,y)}) - \sum_{x\in C}\sum_{i\in I} os^d_{x,i} * log(os^d_{x,i}) \\ &- \sum_{x\in C}\sum_{i\in I} rn^d_{x,i} * log(rn^d_{x,i}) - \sum_{x\in C}\sum_{i\in I} sl^d_{x,i} * log(sl^d_{x,i}) - \sum_{x\in C}\sum_{i\in I} lt^d_{x,i} * log(lt^d_{x,i})\} \end{aligned} \tag{2.1}$$

The objective function of this problem is to maximize the $Entropy$. We estimate the movement probabilities $m^d_{i,j,dist(x,y)}$, which represents the movement probability from sub-population (x, i) to sub-population (y, j) in a week. A sub-population (x, i) is the swine population in the size group i in

the county x. The index variable x and i are used for the originating sub-population, $x = 1, 2, 3...C$ and $i = 1, 2, ...I$. Again, y and j are the index variable for the receiving sub-populations (y, j). The superscript d marks the decision parameters. The parameter $os^d_{x,i}$ represents movement probability from sub-population (x, i) to outside of the state in a week, $rn^d_{x,i}$ is the probability to remain or not-moved in the sub-population (x, i) in a week, $sl^d_{x,i}$ is the probability of pigs being slaughtered for meat from sub-population (x, i) in a week, and $lt^d_{x,i}$ is the probability of pigs being dead or lost in sub-population (x, i) in a week. We divide the distance between counties into five classes: 1) distance $\in [0, 20)$, 2) distance $\in [20, 100)$, 3) distance $\in [100, 200)$, 4) distance $\in [200, 400)$, and 5) distance $\in [400, D_{max}]$. D_{max} is the maximum distance between two counties. $dist(x, y)$ represents the distance class for the distance between county x and y. We divide the distances between all pairs of counties in that way to group them into discrete distance groups. This problem is subject to several linear constraints, which we construct from probability rules, sales data, swine population conservation etc..

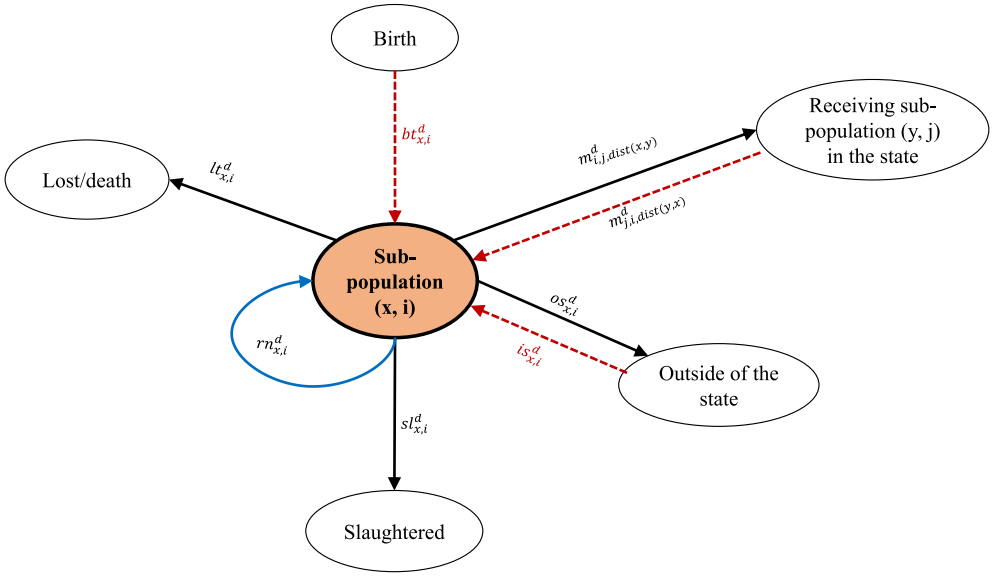


Figure 2.1: **The movement flows of a sub-population** (x, i)**.** Solid black lines represent the outgoing flows from the sub-population, dotted red lines represent the incoming flows into the sub-population, and the blue solid line represents the possibility to stay or not moved. Solid lines (black and blue) form the distributions of the objective function. The probability of each movement are shown with the arrows.

As a pig can move (from the sub-population (x, i) to a sub-population in the state, or outside of

the state, or slaughtered, or death) or it could stay in the sub-population, therefore the summation of these possibilities is equal to one. From the rule of the probability, we can get the following constraint for any sub-population (x, i),

$$\sum_{y \in C} \sum_{j \in I} m^{d}_{i,j,dist(x,y)} + os^{d}_{x,i} + rn^{d}_{x,i} + sl^{d}_{x,i} + lt^{d}_{x,i} = 1 \qquad \forall (x, i) \tag{2.2}$$

The probabilities in Eq 2.2 are considered in the objective function.

There are three types of sales in the system, 1) sales for the movement from sub-population (x, i) to the all sub-populations in the state, 2) sales for the movement to the outside of the state, and 3) sales for slaughter. Constraint for the sales or movement from any county x is,

$$\sum_{i \in I} \sum_{y \in C} \sum_{j \in I} Iv^{r}_{x,i} * m^{d}_{i,j,dist(x,y)} + \sum_{i \in I} Iv^{r}_{x,i} * sl^{d}_{x,i} + \sum_{i \in I} Iv^{r}_{x,i} * os^{d}_{x,i} + ET^{sales}_{x} = \frac{Sales^{r}_{x}}{scaled} \qquad \forall x \tag{2.3}$$

The superscript r indicates published data. The parameter $Iv^{r}_{x,i}$ is the swine inventory in the sub-population (x, i), and $Sales^{r}_{x}$ represents the total sales from county x in a year. The parameter $scaled$ is used to convert the timescale, this parameter allows us to convert the timescale from yearly to weekly basis. ET^{sales}_{x} is the error term for the constraint 2.3.

The constraint for the slaughtered swine is,

$$\sum_{x \in C} \sum_{i \in I} Iv^{r}_{x,i} * sl^{d}_{x,i} + ET^{sl} = \frac{TotalSlaughtered^{r}}{scaled} \tag{2.4}$$

The term $TotalSlaughtered^{r}$ represents the total number of slaughtered in a year in the system, and ET^{sl} is the error term for slaughtered data.

The constraint for the sales to the outside of the state is;

$$\sum_{x \in C} \sum_{i \in I} Iv^{r}_{x,i} * os^{d}_{x,i} + ET^{out} = \frac{TotalOutshipment^{r}}{scaled} \tag{2.5}$$

The term $TotalOutshipment^{r}$ is the total sales to the outside of the state in a year, and ET^{out} is the error term for outshipment.

The constraint for the inshipments from the outside of the state is;

$$\sum_{x \in C} \sum_{i \in I} Iv^{r}_{x,i} * is^{d}_{x,i} + ET^{in} = \frac{TotalInshipment^{r}}{scaled} \tag{2.6}$$

The parameter $is^{d}_{x,i}$ indicates the inshipment probability in a week from outside of the state to the sub-population (x, i), $TotalInshipment^{r}$ is the inshipment from outside in a year in the system, and ET^{in} is the error term for inshipment.

The constraint for the death or lost is,

$$\sum_{x \in C} \sum_{i \in I} Iv^{r}_{x,i} * lt^{d}_{x,i} + ET^{lt} = \frac{TotalLost^{r}}{scaled} \tag{2.7}$$

The term $TotalLost^{r}$ represents the total number of death or lost in a year from the system, and ET^{lt} is the error term for this constraint.

We assume that the population or inventory size of a sub-population remain constant throughout the year. Therefore, in a sub-population, the summation of the outgoing flows from the sub-population (solid black lines in Fig A.2) is equal to the summation of the incoming flows into the sub-population (dotted red lines in Fig A.2). Constraints for the population conservation are,

$$\begin{aligned} & Iv^{r}_{x,i} * [\sum_{y \in C} \sum_{j \in I} m^{d}_{i,j,dist(x,y)}] + Iv^{r}_{x,i} * sl^{d}_{x,i} + Iv^{r}_{x,i} * lt^{d}_{x,i} + Iv^{r}_{x,i} * os^{d}_{x,i} \\ = & \sum_{y \in C} \sum_{j \in I} Iv^{r}_{y,j} * m^{d}_{j,i,dist(y,x)} + Iv^{d}_{x,i,b} * bt^{d}_{x,i} + Iv^{r}_{x,i} * is^{d}_{x,i} + ET^{pop}_{x,i} \qquad \forall (x, i) \end{aligned} \tag{2.8}$$

Here, $Iv^{d}_{x,i,b}$ represents the breeding population, $bt^{d}_{x,i}$ is the probability of birth in the sub-population (x, i) in a week, and $ET^{pop}_{x,i}$ is the error term. The left side of the Eq. 2.8 is the summation of the outgoing flows from sub-population (x, i) and the right side is the summation of the incoming flows into the sub-population (x, i). The range for $bt^{d}_{x,i}$ is $(7\times9)/115 - (7\times12)/112\ week^{-1}$, as time period for gestation is 112-115 days and average litter rate is 9-12 [61]. The range for $sl^{d}_{x,i}$ was chosen based on the lifespan of market pigs in the US, which is about 25 to 28 weeks.

Constraint for the errors is,

$$\sum_{x \in C} |ET_x^{sales}| + |ET^{sl}| + |ET^{in}| + |ET^{out}| + ET^{lt} + \sum_{x \in C} \sum_{i \in I} |ET_{x,i}^{pop}| \leq R_c * TotalPopulation^r \quad (2.9)$$

The left side of Eq. 2.9 represents the summation of all the errors in the optimization problem. Here, R_c is a proportional constant, and $TotalPopulation^r$ is the total swine population in the system. The inequality (Eq. 2.9) states that the total error in the convex optimization problem should be less than equal to a fraction R_c of the $TotalPopulation^r$. The value of R_c is calculated by using trial and error with an objective to minimize the total error.

Convex cost function (Eq. 2.1) and constraints (Eq. 2.2-2.9) constitute our optimization linear problem. The objective of this estimation problem is to maximize the entropy of the distributions of the objective function of all sub-populations. The performance of entropy measures is sensitive to different factors [67]. Maximum entropy methods can predict accurately given a prior knowledge. However, maximum entropy methods can perform poorly if the prior knowledge is insufficient or inaccurate or contains biases [68]. In our estimation problem, published USDA-NASS data are used as the prior knowledge, and the data was sufficient to solve the formulated convex optimization problem. Maximum entropy methods can also perform poorly if the system changes very rapidly [68], which is not our case.

2.4 Network development

We develop a network using the movement parameters which are obtained using the maximum entropy optimization. The network development is done in two stages: 1) setup of the population in each farm and 2) setup of the movement links between farms.

In order to generate the network, first, we need the farm-level estimates of the pig population. The USDA-NASS data only provide the number of farms in a size range and the number of total pigs in that range in a county. Recorded data on the number of pigs in a farm are generally not available in the US (with the exception of a few counties). To allocate the pig population, we generate random numbers for every farm in a size group i within a county x with the following

constraints:

a) The random numbers fall in the range of the corresponding group i.

b) The sum of all generated numbers is equal to the total number of pigs in that sub-population (x, i).

The procedure to establish the movement links between farms is inspired by the random network model [69]. Our movement network for pig farms is represented as (V, E, W). The term V denotes the set of nodes, the term E represents the set of links or connections among individual nodes, and W denotes the weight of each link. To generate the movement network among farms, we use the following procedures:

Step 1 For each pig p_1 in a sub-population (x, i), we generate a random number *rand* from the uniform distribution $U(0, 1)$ for sub-population (y, j), $y = 1, 2, 3, ...C$, and $j = 1, 2, 3, ...I$. Here, C is the number of counties in the system and I is the number of size groups.

Step 2 If $rand <= m^{d}_{i,j,dist(x,y)}$, a link is created from pig p_1 to another pig p_2. Pig p_2 is picked randomly from the sub-population (y, j).

Step 3 If there is no link from the parent farm f_1 of pig p_1 to the parent farm f_2 of pig p_2, we create a link $flink$ from f_1 to f_2. Otherwise, if a link already exists, we increase its weight by 1.

Step 4 For each sub-population (x, i), we repeat Steps 1-3.

This process produces a directed weighted network at the farm-level. Links or connections among farms represent swine movement. The weight of a link represents the volume of movements occurring from one farm to another.

2.5 Network analysis

To capture the particular features of the developed network, we compute the following network analysis metrics: node strength, betweenness, eigenvector, clustering coefficient, and average

shortest path [24, 38, 69]. Centrality measures can help us determine the most important or central nodes in a network.

The **node strength-centrality** measure is the strength of the nodes or sum of the weights of the edges connected to it [70]. In a directed network, the nodes have two types of vertex-strength centralities: 1) in-strength InS, and 2) out-strength OuS.

$$InS(k) = \sum_{l \in NB(k)} w_{lk} \tag{2.10}$$

$$OuS(k) = \sum_{l \in NB(k)} w_{kl} \tag{2.11}$$

Here, w_{lk} is the connection strength of the edge/link from node l to node k, $NB(k)$ is the set of the neighbors of node k. Vertex strength can be illuminating in the investigation of epidemic processes. A high in-strength node has a high risk of receiving an infection. On the other hand, a high out-strength node is influential over the network, as such a node can infect many more nodes.

The **betweenness centrality** measure suggests which nodes are important in the connection flow or act as bridges in the network. Betweenness centrality of a node measures how many shortest paths between different pairs of nodes go through that particular node. The shortest path between two nodes is the path with the fewest number of connections. Nodes with high betweenness centrality have high control over movement flow (here, concerning flow of swine) in the network. Removal of such nodes can effectively reduce connectivity in the network. Knowledge of these nodes can be useful in controlling outbreaks [71]. Let, p_{st} be the number of shortest paths from $s \in N$ to $t \in N$. We denote, $p_{st}(k)$ to be the number of shortest paths from s to t, that includes node k somewhere in between. The betweenness centrality of a node k is defined [72] as:

$$B(k) = \sum_{s \neq k \neq t \in N} \frac{p_{st}(k)}{p_{st}} \tag{2.12}$$

Eigenvector centrality is an extension of the degree/strength centrality. In the eigenvector centrality measure, the centrality of a node is proportional to the sum of the centralities of its

neighbors.

$$e(k) = \lambda_1^{-1} * \sum_{l \in NB(k)} e(l) \tag{2.13}$$

Here, $e(k)$ is the eigenvector centrality of the node k, and λ_1 is the largest eigenvalue of the adjacency matrix $[a_{kl}]$ of the network. Eigenvector centrality of a node can be large if either it has many neighbors or it has important neighbors. Nodes with high eigenvector centralities have high probabilities of becoming infected [16, 73].

The **clustering coefficient** measures local group cohesiveness. The clustering coefficient *Cc(k)* for a node k is the ratio of the number of edges among the neighbors of k and the maximum possible number of such edges (for the fully-connected network formed by the neighbors of node k). If neighboring nodes of node k has c_k connections among them then clustering coefficient can be defined as [24]:

$$Cc(k) = \frac{c_k}{|NB(k)|(|NB(k)| - 1)/2} \tag{2.14}$$

The **average shortest path** is the average of the shortest path length between all pairs of nodes in the network.

2.6 Farm-level movement network for Iowa

2.6.1 Movement probability estimation

In this research, we solve a convex optimization problem to estimate the swine movement probabilities by using the maximum entropy approach for Iowa. We utilized the AIMMS modeling system [74] of Paragon Decision Technology to solve this convex optimization problem. The time-scale of our estimation problem is weekly, which we controlled it by using $scaled = 52 weeks/year$. The boundary of error limit in our system is 5.45% of total swine population in Iowa (R_c=5.45%). The estimated probabilities are given in Table 2.1. This table shows swine movement probabilities between size groups for five different distance ranges. The highest movement probability is from size7 to size7 sub-population when the distance between them is less than 20km. We divide seven size groups into three categories; size: 1-3(small farms), 4-5(medium farms), and 6-7(large farms).

From Table 2.1, we can notice that the movement probabilities from large farms to small farms are small and vice versa.

2.6.2 Network description

We generate a swine movement network for the central agricultural district of Iowa. It has 12 counties: Boone, Dallas, Grundy, Hamilton, Hardin, Jasper, Marshall, Polk, Poweshiek, Story, Tama, and Webster. The total number of farms in those 12 counties is 641, while the net pig population is 2,600,888, which is 12.71% of the total pig population in Iowa. Grundy, Hamilton, Hardin, Jasper, Marshall, and Webster Counties are within the America's top 100 pork producer counties. Among these, Hardin County is in the 9^{th} position. The descriptions of pig inventories for the above-mentioned counties are provided in the supplementary material Dataset 1.

For these 12 counties, we have developed a movement network (V, E, W), which is shown in Fig 2.2. This network is a realization based on the movement probabilities from Table 2.1. For the network, $|V| = 641$ and $|E| = 22,461$, the description of the nodes, and the adjacency list for this network is provided in the supplementary material Dataset 2 and 3. In Fig 2.2, this network has seven types of nodes representing the seven size groups. A description of size groups is presented in Table A.1. The largest group is the size7, contains 393 nodes which are presented by light blue. There are 17484 edges among the nodes of this group (67.41% of total edges).

2.6.3 Network analysis

The clustering coefficient of the full network is 0.363, the diameter of the network is 7, and the average shortest path length is 2.598. A summary of various centrality measures for the network is provided in Table 2.3. Node-strength, betweenness, eigenvector and clustering coefficient centrality for seven size groups are presented here. In-strength, out-strength, betweenness, and eigenvector centralities were calculated from the overall network. Clustering coefficients in Table 2.3 were calculated for networks of the same size group (any node and its neighbors are in the same size group). We used the open source package Gephi to visualize and analyze the network [75]. For visualization, we used the Fruchterman Reingold layout [76].

Table 2.1: Estimated swine movement probabilities $m_{i,j,dist(x,y)} \times 10^3$ from maximum entropy approach.

		Destination						
		Size1	Size2	Size3	Size4	Size5	Size6	Size7
Distance < 20*km*								
	size1	1.4899	1.3587	1.3890	1.4007	1.4543	1.4641	1.5083
	size2	1.3989	1.5080	1.3755	1.4129	1.4393	1.4611	1.5112
	size3	1.2826	1.1726	1.8054	1.4979	1.5580	1.6066	1.6264
Source	size4	1.0582	1.1064	1.4199	2.3913	1.7695	1.9519	2.1038
	size5	0	0	0	1.7460	7.1795	6.0844	5.3446
	size6	0	0	0	2.5308	8.7793	14.3449	8.8213
	size7	0	0	0	0	0	0	11.7828
20*km* < Distance < 100*km*								
	size1	1.3334	1.3028	1.3834	1.4076	1.4403	1.4511	1.4972
	size2	1.3373	1.2961	1.3767	1.4114	1.4375	1.4463	1.4987
	size3	1.2407	1.1528	1.3516	1.4039	1.5402	1.5589	1.6340
Source	size4	1.0077	0.7768	1.2906	1.3337	1.7005	1.7403	1.9707
	size5	0	0	0	0.5768	2.4553	3.4121	4.4916
	size6	0	0	0	0	2.0213	4.0961	6.3753
	size7	0	0	0	0	0	0	0
100*km* < Distance < 200*km*								
	size1	1.3211	1.2904	1.3840	1.3943	1.4421	1.4449	1.5056
	size2	1.3261	1.3009	1.3899	1.3914	1.4372	1.4392	1.4987
	size3	1.2350	1.1626	1.3534	1.3966	1.4823	1.5003	1.6312
Source	size4	0.9633	0.7990	1.3194	1.3922	1.6203	1.6701	1.9975
	size5	0	0	0	0.2870	2.0726	2.2576	4.5535
	size6	0	0	0	0	0.7503	1.2075	6.5958
	size7	0	0	0	0	0	0	0
200*km* < Distance < 400*km*								
	size1	1.3092	1.2929	1.3708	1.3906	1.4435	1.4587	1.5156
	size2	1.3101	1.2912	1.3705	1.3919	1.4453	1.4608	1.5130
	size3	1.1890	1.1582	1.3361	1.3725	1.4957	1.5190	1.6690
Source	size4	0.9148	0.8430	1.2363	1.3534	1.6271	1.6868	2.0233
	size5	0	0	0	0.0996	1.9382	2.2667	4.8693
	size6	0	0	0	0	0.1753	0.7087	7.3607
	size7	0	0	0	0	0	0	0
Distance > 400*km*								
	size1	1.2644	1.2818	1.3040	1.4093	1.4522	1.5169	1.5613
	size2	1.2915	1.2876	1.3032	1.4002	1.4492	1.5108	1.5422
	size3	1.1489	1.1554	1.1864	1.4614	1.4731	1.6829	1.7441
Source	size4	0.9891	0.8387	0.9770	1.4179	1.6056	1.9855	2.0836
	size5	0	0	0	0.1091	0.8917	3.9986	4.4802
	size6	0	0	0	0	0	3.3953	5.4755
	size7	0	0	0	0	0	0	0.0019

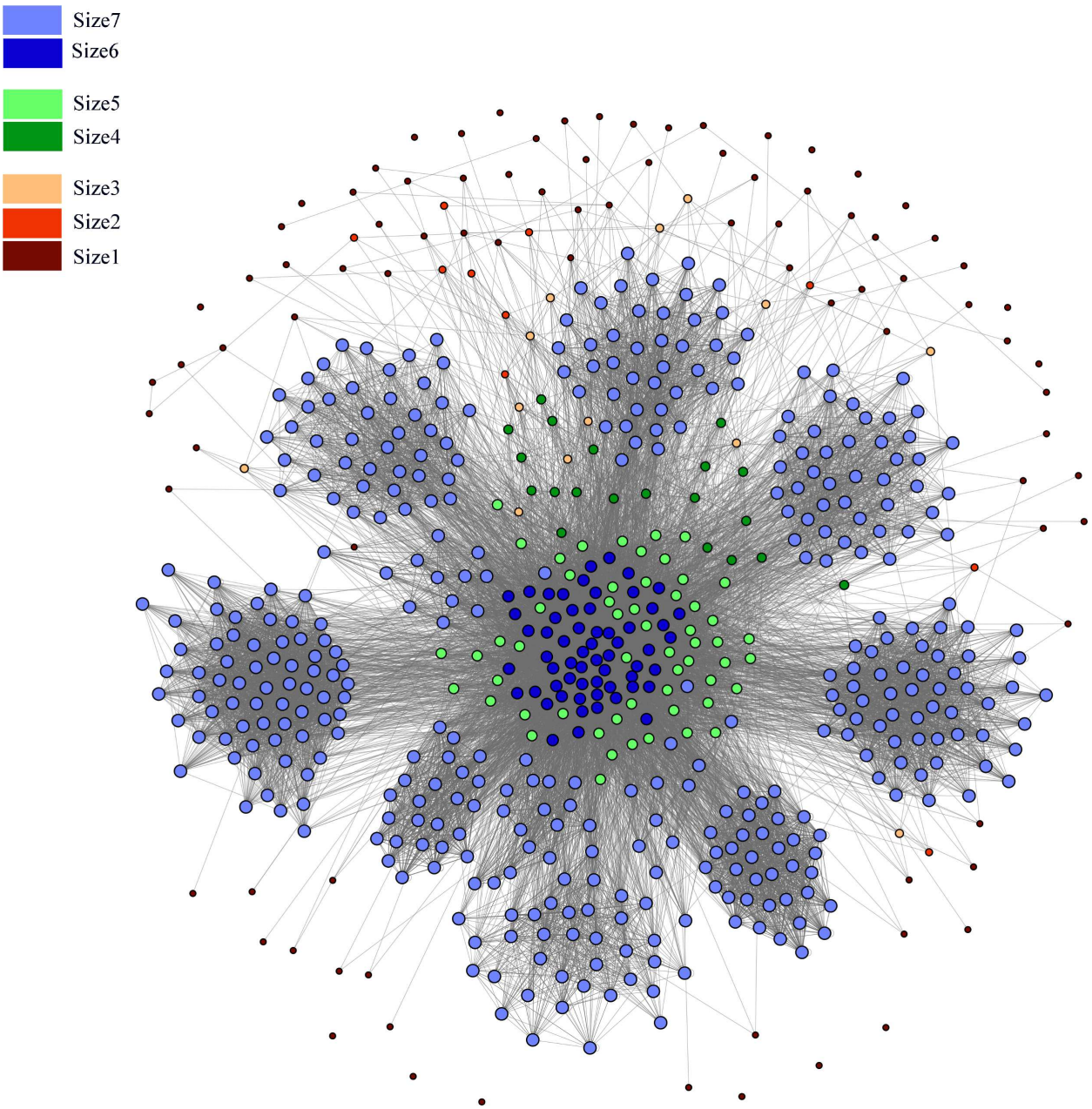


Figure 2.2: **Movement network for the pig population at the farm-level.** Different colors represent different size groups. Farms are divided into 7 size groups, size: 1-3(small farms), 4-5(medium farms), and 6-7(large farms).

From the node-strength centrality measures, we observe that the average node-strength is positively correlated with the size groups. Larger size groups have higher average node-strengths. Consequently, size7 has the highest average node-strength (Table 2.3). The node-strength distribution is provided in Fig A.7. In the network, only a few nodes have high strength and most of the

Table 2.2: A summary of the size groups in the network.

Group	No. of nodes	% of the total nodes	No. of edges in a group	% of the total edges
size1	89	13.88%	15	0.07%
size2	10	1.56%	2	0.01%
size3	13	2.03%	9	0.04%
size4	20	3.12%	50	0.22%
size5	56	8.74%	678	3.02%
size6	60	9.36%	1666	7.42%
size7	393	61.31%	12506	55.68%

nodes have low strength. This characteristic is similar to the power-law distribution. The range of in-strength is $0 - 1426$. About 90.95% of the total nodes have in-strengths less than 285, which is merely the first 20% of the in-strength range. The range for out-strength is $0 - 1372$. About 91.11% of the total nodes have out-strengths less than 274, which is within the first 20% of the range of out-strength values. The correlation coefficient between in-strength and out-strength is 0.9523, which is an indication of strong correlation.

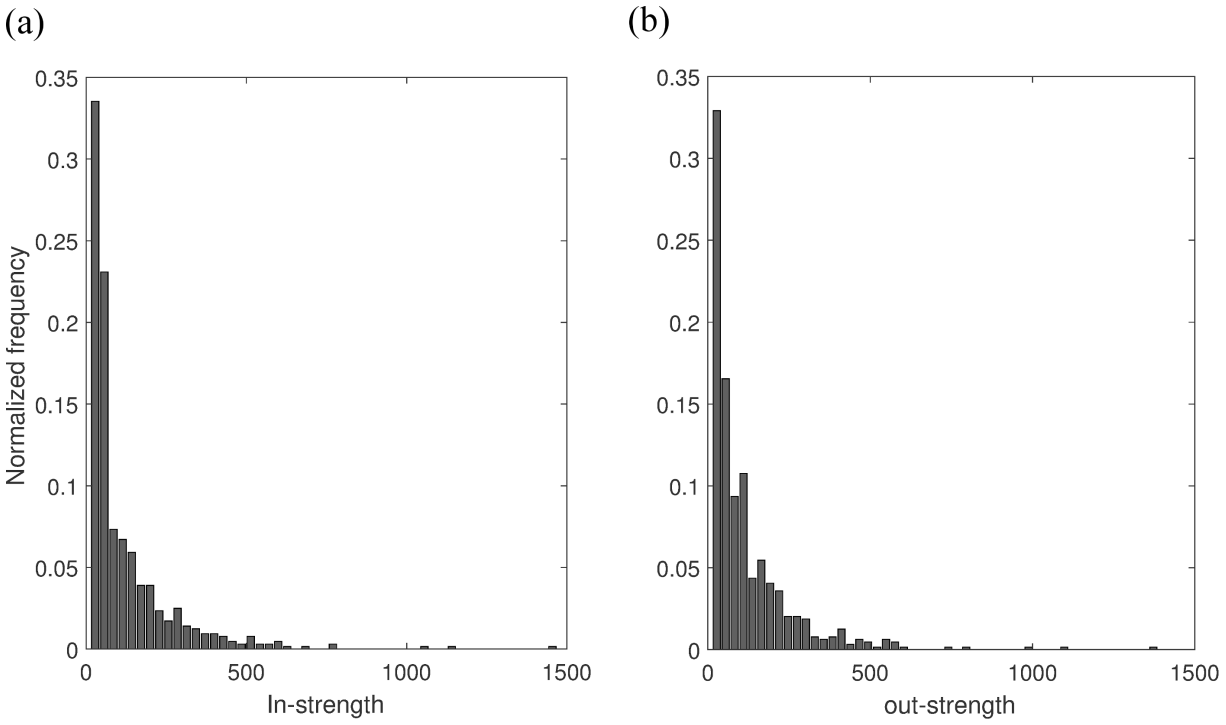


Figure 2.3: **Node strength distribution of the directed network**. (a) in-strength, (b) out-strength

The betweenness centrality is positively correlated with size groups until group6, after which

Table 2.3: A summary of centrality measures for different size groups in the network.

	Size1	Size2	Size3	Size4	Size5	Size6	Size7
In-strength							
mean	1.292	4.700	6.846	16.050	44.304	63.400	151.891
median	1.000	4.000	5.000	15.000	31.500	44.000	100.000
(95% CI)	(0.902, 1.683)	(2.620, 6.780)	(4.214, 9.478)	(11.781, 20.319)	(33.070, 55.537)	(48.566, 78.234)	(135.376, 168.406)
range	(0, 8)	(1, 9)	(1, 17)	(5, 42)	(12, 267)	(18, 347)	(11, 1426)
Out-strength							
mean	1.214	4.500	11.385	22.200	55.054	138.450	140.461
median	1.000	3.500	9.000	18.500	53.000	109.5000	90.000
(95% CI)	(0.935, 1.491)	(1.613, 7.386)	(6.746, 16.023)	(16.830, 27.569)	(48.889, 61.217)	(123.497, 153.403)	(122.477, 154.444)
range	(0, 5)	(0, 14)	(2, 26)	(10, 50)	(21, 118)	(66, 282)	(7, 1372)
Betweenness							
mean	36.140	386.258	858.157	1531.4	814.294	2390.600	244.137
median	0	86.087	905.169	1289.900	661.0194	2026.000	132.247
(95% CI)	(10.639, 61.642)	(4.551, 767.964)	(413.300, 1303.000)	(1127.400, 1935.300)	(634.840, 993.748)	(1738.500, 3042.600)	(183.130, 305.143)
range	(0, 699.662)	(0, 1237.000)	(14.605, 2388.400)	(228.138, 3189.900)	(48.185, 2715.600)	(324.236, 15229.000)	(0.256, 9932.100)
Eigenvector							
mean	.00086	0.0032	0.0058	0.0326	0.1072	0.1522	0.2381
median	.00035	0.0030	0.0038	0.0279	0.0854	0.1263	0.1690
(95% CI)	(0.0006, 0.0012)	(0.0020, 0.0044)	(0.0033, 0.0083)	(0.0235, 0.0417)	(0.0899, 0.1245)	(0.1225, 0.1819)	(0.2174, 0.2588)
range	(0, 0.0061)	(0.0011, 0.0064)	(.00043, 0.0141)	(0.0100, 0.0726)	(0.0281, 0.3391)	(0.0493, 0.6565)	(0.0328, 1)
Clustering coefficient							
mean	0	0	0	0.124	0.264	0.449	0.755

farms in the group7 have lower betweenness. The farms in group6 have the highest average betweenness. The distribution of betweenness centrality measure is given in Fig A.8. Most of the farms have low betweenness. Few farms act as hubs in the network which have high betweenness. The range for betweenness is 0-15229. We divide the nodes into three groups, 1) low-betweenness (0-50), 2) medium-betweenness (51-500), and 3) high-betweenness (> 500). These three groups contain 183, 302, and 156 nodes respectively. These three groups are illustrated in Fig 2.5. In the low-betweenness group majority of the nodes are from small size groups, in the medium-betweenness group most of the nodes are from group7, and in the high-betweenness group, most of the nodes are from group6.

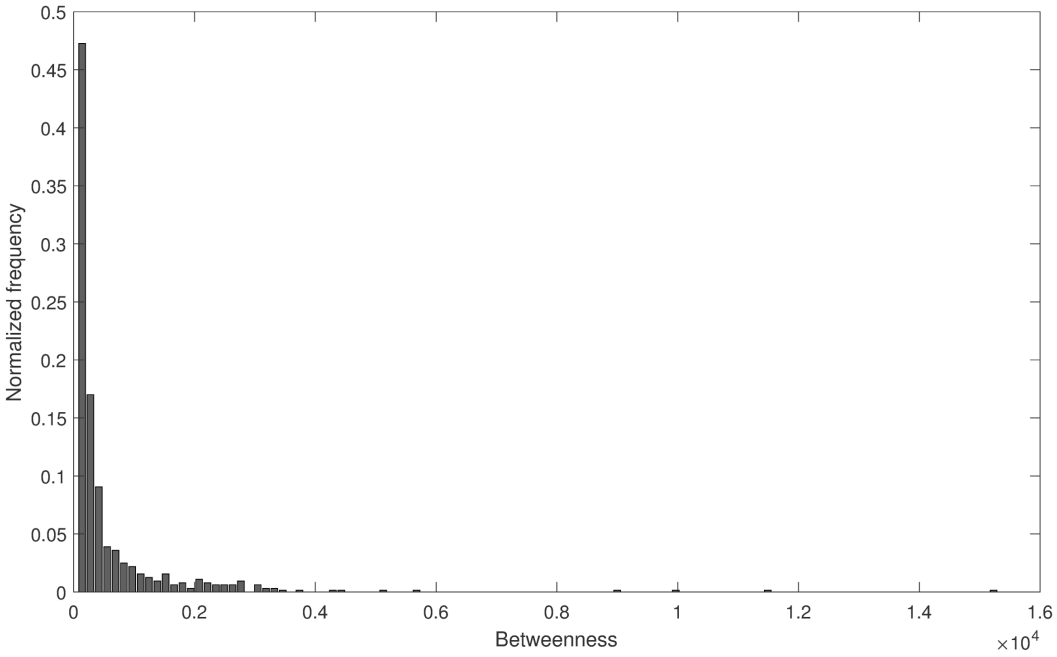


Figure 2.4: **Betweenness distribution of the network**.

The mean eigenvector centrality is positively correlated with the size groups. Larger size groups have higher eigenvector centralities (Table 2.3). We have divided the nodes (farms) into three groups: 1) low-eigenvector central nodes (0-0.1), 2) medium-eigenvector central nodes (0.11-0.3), and 3) high-eigenvector central nodes (0.31-1). The low-eigenvector central group consists of 298 nodes, the medium group consists of 233 nodes, and the high group contains the rest of the nodes. The network for different eigenvector groups is presented in Fig 2.6. Clustering coefficient

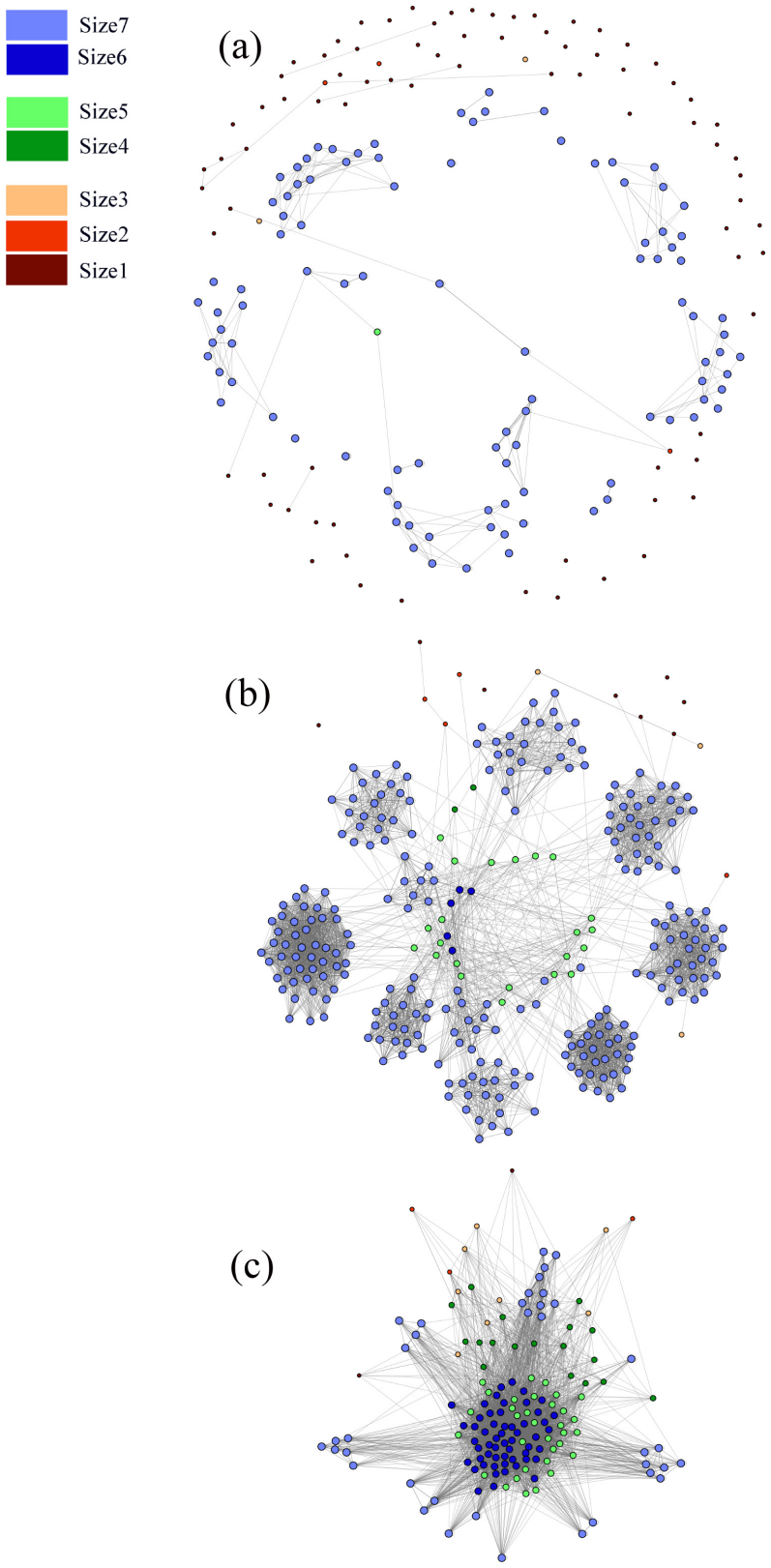


Figure 2.5: **Node groups according to betweenness**. a) nodes with low-betweenness, b) nodes with medium-betweenness, and c) nodes with high-betweenness. The connections among visible nodes are presented here.

for group size 7 is 0.755, which is quite high. The nodes from this group form several clusters, which are quite visible in Fig 2.2 and Fig 2.6.

In the network, the importance of links is another useful topic to study [52]. From the link strength or weight distribution, we can see that the majority of the links have a low weight however very few links have high weight (Fig 2.7). A link with high-weight represents a high volume swine movement. For a susceptible farm, an infected neighbor connected by a high-strength-link is riskier than an infected neighbor connected by a low-strength-link.

2.7 Summary

In this study, we have three objectives: 1) we compute optimal estimates swine movement probabilities among counties from the aggregated data of USDA-NASS, 2) we develop a realization of the network from the estimated probabilities, and 3) we analyze the developed network with different network analysis metrics.

Animal movement has been one of the major causes of diseases propagation among farms for several outbreaks in the US swine industry. A better understanding of the swine movement network can increase the feasibility of planning effective mitigation strategies that can reduce the risk of disease dispersal. There is no mandatory animal movement tracking system in the US due to the industry preference for privacy in the swine business. We have estimated the movements among different swine sub-populations using a convex optimization problem, have formulated according to the USDA-NASS data. The discrepancy from our optimization problem is about 5.45% of the total swine population, which is slightly higher than that of a similar work on cattle movement probability estimation [29] due to a greater amount of data available for cattle. Our estimation can be improved if more data are available. The additional data that would improve the results most is the type of swine operations (for example, nursery, farrow-to-feeder, farrow-to-wean, farrow-to-finish, finish only etc.) at the county level. The USDA-NASS department can collect and publish this information in future reports, as this additional data would not hamper the anonymity of the Census of Agriculture yet greatly improve movement estimations.

The network development algorithm can provide us a realization of the network from the esti-

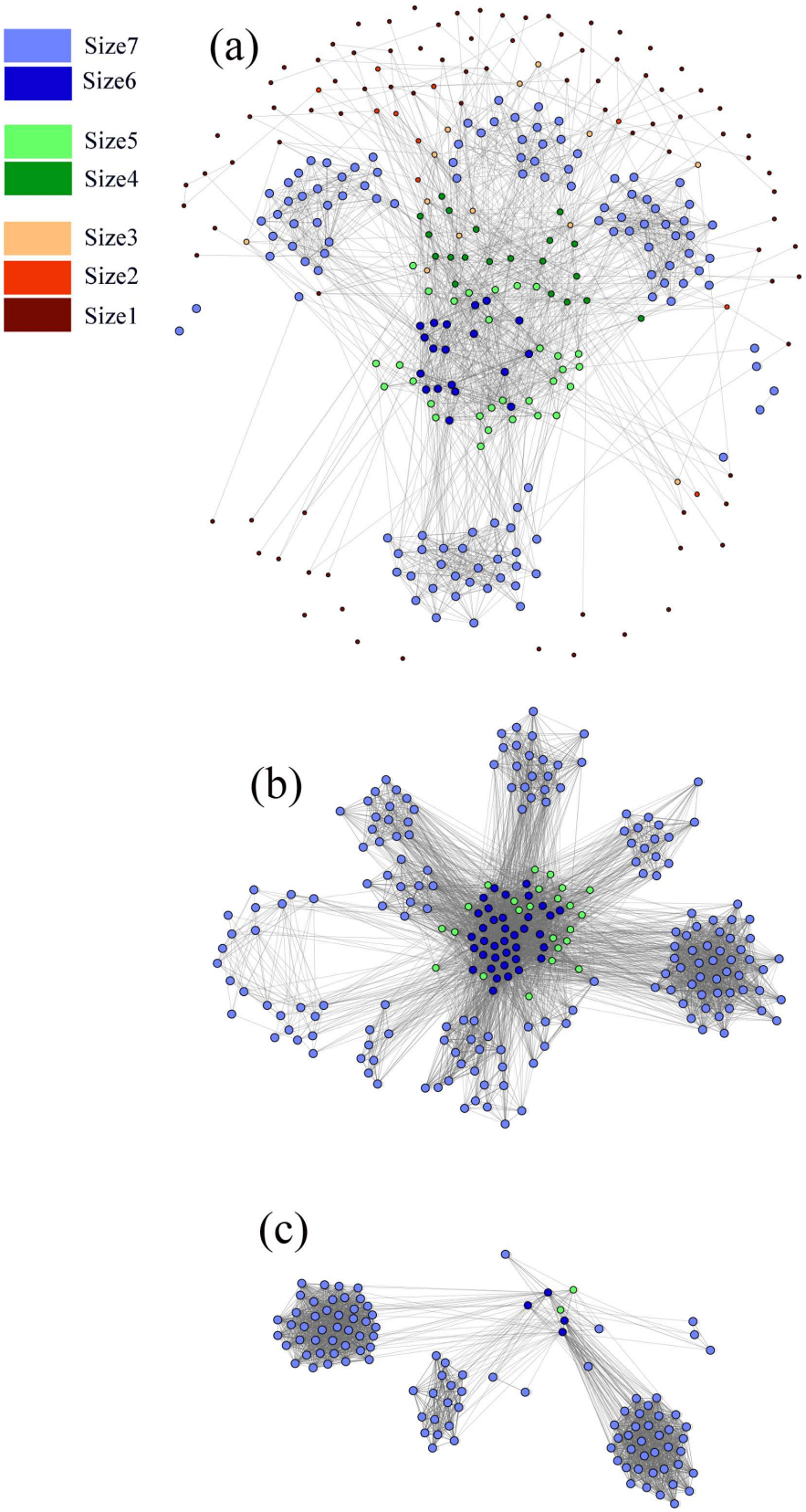


Figure 2.6: **Node groups according to eigenvector centrality**, a) low-eigenvector central nodes, b) medium-eigenvector central nodes, and c) high-eigenvector central nodes. The connections among visible nodes are presented here.

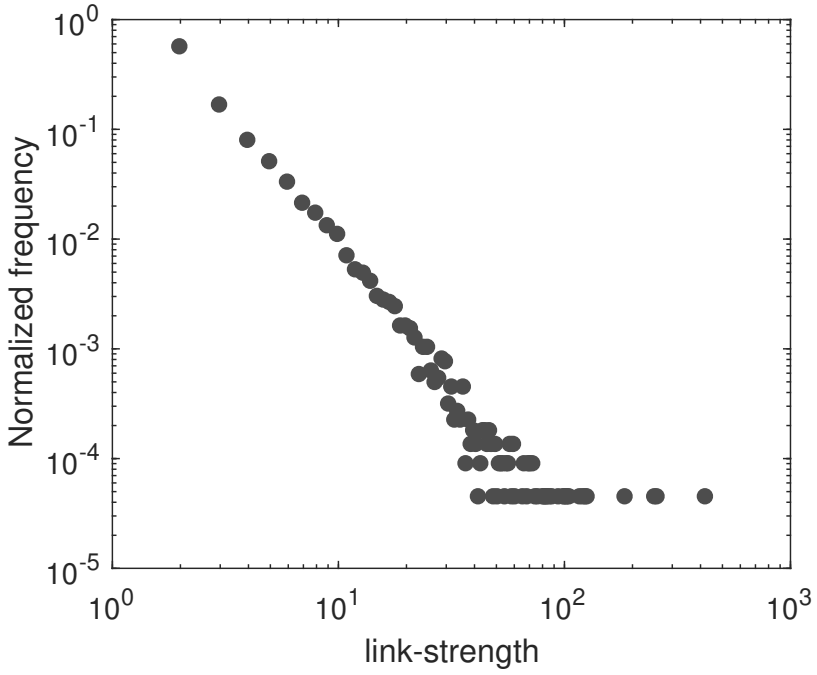


Figure 2.7: **Link-strength or connection-weight distribution of the network**. Log-log scale has used for better visualization.

mated movement probabilities. The generated swine movement network was well connected with a giant component containing 95.94% of the farms. The implication of this high connectivity is that the swine industry may be vulnerable to infectious diseases. All the disconnected farms were smaller farms (inventory size less than 100) where most of them produce meat for their own consumption (60.5% of all small swine farms) [77]. In addition to that, most of these small farms are engaged in all of the phases of swine production (farrow-to-finish producers) [78]. On the other hand, larger farms have more connections among them. One possible reason could be that most of the large farms are specialized in a single production phase to increase productivity [79, 80]. Consequently, pig shipments are very frequent among them.

We use centrality measures to understand the characteristics of the movement network. From the analysis of the node-strength centrality measure, we notice that many nodes in the network have low node-strength however very few nodes have high node-strength, who work as hubs in the network. The node-strength distribution of the network is similar to that of scale-free networks (Fig A.7). Compared to a random network, epidemics can disperse faster in a scale-free network.

In addition to that, scale-free networks have lower epidemic threshold than comparable random networks [81]. This information could be useful because targeted vaccination/node-removal is more effective in scale-free structures than random vaccination [82]. The vaccination, or culling, or quarantine of the hubs (farms with high node-strength) can be crucial to control an epidemic.

If we analyze the average shortest path length and the clustering coefficient of the overall network, we see evidence of the small-world phenomenon in the network. The average path length was similar and clustering coefficient was more than six times larger compared to the similar properties of the equivalent Erdos-Renyi random network [83], which satisfy the sufficient conditions for small-world properties of the network [84]. The US swine movement network structure is quite vulnerable to any disease propagation because of its small-world nature. This result is similar to other studies as well [55–57]. This network has high local clustering. Size7 group (larger operations: headcount is more than 1000) has the highest amount of local clustering (Fig 2.2 and Fig 2.6). Therefore, large operations are highly interconnected, making them more vulnerable to outbreaks. Moreover, the structure of the US swine industry has been changing over several years. The number of large operations is increasing, where most of them specialize in one particular phase of production. These changes are increasing the risk for disease outbreaks in the swine industry.

The correlation between in-strength (incoming movements) and out-strength (outgoing movements) is strong. The nodes with high out-strength values also have high in-strength values. This is an important indicator as the nodes with a high risk of receiving infection are also highly capable of transmitting them.

Although the group size7 (largest operations) has the highest values of node-strength, clustering coefficient, and eigenvector centralities it is not necessarily highest in terms of the betweenness centrality measure. We found that group size6 has the highest betweenness centrality values (Table 2.3). The groups size4 and size5 also show high betweenness. The above-mentioned properties indicate that the group size7 forms various clusters in the network, where the operations are highly connected. The operations of medium size, however, maintain the connectivity among the clusters of the largest group. Hence, these medium size operations play a key role in the system. During an epidemic, it is possible to use these high betweenness farms to disconnect the movement network and confine the disease in a smaller part of the network.

We make several assumptions to simplify our model as all necessary data are not available. We assume that the inventory size of the operations is constant on a year-to-year basis. We also consider that movement flows are the same throughout the year because of the resolution limitation of the available data. However, movement flows can be different from one season to another season. The movement flows also can be sensitive to other factors, for example, production technology, business strategy, and food availability. However, we do not have specific knowledge about these factors at this point and inclusion of too many unknown factors increases the complexity and uncertainty of the estimation problem given the limited data. Our estimation steps can be easily adapted by adding more constraints when more data are available.

One immediate use of this network could be the investigation of the stochastic epidemic processes [1, 11, 85–87]. This kind of study can help us understand the underlying mechanisms and threshold conditions of epidemic processes for various swine diseases including porcine reproductive and respiratory syndrome (PRRS), classical swine fever (CSF), African swine fever (ASF) and many more.

In summary, we present a maximum entropy approach to estimate the swine movement network from aggregated anonymous census data. This method can be used to estimate movement probabilities of other farm animals too for various locations.

2.8 Data availability

The dataset used to perform this research is available from https://quickstats.nass.usda.gov/. The authors are willing to provide further details upon request.

Chapter 3

Contact tracing evaluation for COVID-19 transmission in the different movement levels of a rural college town in the USA[1]

Contact tracing can play a key role in controlling human-to-human transmission of a highly contagious disease such as COVID-19. We investigate the benefits and costs of contact tracing in the COVID-19 transmission. We estimate two unknown epidemic model parameters (basic reproductive number R_0 and confirmed rate δ_2) using confirmed case data. We model contact tracing in a two-layer network model. The two-layer network comprises the contact network in the first layer and the tracing network in the second layer. In terms of benefits, simulation results show that increasing the fraction of traced contacts decreases the size of the epidemic. For example, tracing 25% of the contacts is enough for any reopening scenario to reduce the number of confirmed cases by half. Considering the act of quarantining susceptible households as the contact tracing cost, we have observed an interesting phenomenon. The number of quarantined susceptible people increases with tracing because each individual confirmed case is mentioning more contacts. However, after reaching a maximum point, the number of quarantined susceptible people starts to decrease with the increase of tracing because the increment of the mentioned contacts is balanced

[1]This chapter is a slightly modified version of our published article [39], Copyright ©2021, Scientific Reports.

by a reduced number of confirmed cases. This research aims to assess the effectiveness of contact tracing for the containment of COVID-19 transmission in the different movement levels of a rural college town in the USA. Our research model is designed to be flexible and can be used in other geographic locations.

3.1 Background

COVID-19 has affected the lives of billions of people in 2019-2020. The COVID-19 disease is caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) and has caused a global health emergency. The world health organization (WHO) declared it a Public Health Emergency of International Concern on January 30, 2020 [88]. The number of confirmed reported cases by SARS-CoV-2 has been rising. On May 31, 2020, worldwide there were $5,939,234$ laboratory-confirmed cases with $367,255$ deaths [89].

Many countries issued a pandemic lockdown to slow down the COVID-19 transmission. In the United States, a "Stay-At-Home" order was issued in many states. However, those pandemic lockdowns have a massive impact on the economy. All the States of the USA started reopening gradually from early May. Understanding the impact of mitigation strategies on the spreading dynamic of COVID-19 during the reopening phase of the USA is essential. In this work, we assess the impact of contact tracing using an individual-based network model under four reopening scenarios: 25% reopening, 50% reopening, 75% reopening, and 100% reopening (no restriction).

Individual-based contact-network models are a powerful tool to model COVID-19 dispersal due to COVID-19's person-to-person transmission nature. In this work, we develop an individual-based network model for a college town, Manhattan, KS, where households represent nodes of the network. We select Manhattan, KS, as our study area since it is a typical college town in a rural region of Kansas, the home of Kansas State University. There are $20,439$ occupied households in Manhattan, KS, according to census 2018 [90]. The connections between two individual households represent the contact probabilities between the members of the households. To develop the contact network, we consider age-stratification and use Google COVID-19 community mobility reports [91]. The individual-based approach provides the flexibility to observe the local dynamic at the

individual level. It also allows us to include a mitigation strategy in the model at the individual level, such as contact tracing.

Designing an epidemic model for COVID-19 is challenging, as many epidemic features of the disease are yet to be investigated, such as, for example, the transmission rate, the pre-symptomatic transmission rate, and the percentage of the asymptomatic population. These uncertain characteristics make epidemic modeling challenging as the outcomes of the model are sensitive to the assumption made on the uncertainties. Therefore, we use a simple epidemic model with five compartments –susceptible-exposed-infected-confirmed-removed (SEICR)– capable of imitating the COVID-19 transmission and flexible enough to cope with new information. This model has only two unknown parameters: the basic reproductive number R_0 and the confirmed case rate or reporting rate δ_2. An analytical/numerical approach to the computation of R_0 can be found in Barril et al. [92] and Breda et al.[93], respectively. We use COVID-19 cases from March 25, 2020, to May 4, 2020, in Manhattan, KS as data, and estimate the unknown parameters. We use this period to estimate R_0 as there was no reopening in Manhattan, KS; therefore, the contact network was the same through the whole time. The other parameters are taken from the literature. In the COVID-19 transmission, there are pre-symptomatic and asymptomatic cases that may not show any sign of illness [94]. Besides, there is a strong possibility that infected cases not detected exist. In our epidemic model, we have considered those unreported cases. We assume that a confirmed COVID-19 patient cannot transmit the disease anymore except in his/her household.

Since a vaccine is not available at the time of this writing in May 2020 for COVID-19, contact tracing is a key mitigation strategy to control the COVID-19 transmission. Contact tracing is a mitigation strategy that aims at identifying people who may have come into contact with a patient. This mitigation strategy prevents further transmission by quarantine of exposed people. The public health personnel have used contact tracing as a tool to control disease dispersal for a long time [95]. We implement two approaches of the contact tracing strategy through a two-layer network model with two modified SEICR epidemic models. In the first contact tracing approach, we consider all the traced contacts of a confirmed case will be quarantined, which follows the CDC contact tracing guidance for COVID-19 (October 21, 2020) [95]. In the second contact tracing approach, we consider only the tested positive traced contacts of a confirmed case will be isolated. We propose

two quarantine approach to compare their effectiveness. This research finds that quarantine all the traced contacts is always effective than quarantine only test positive traced contacts. The feasibility of contact tracing to control COVID-19 transmission was analyzed using a branching process stochastic simulation for three basic reproductive numbers $R_0 = 1.5, 2.5$, and 3.5 [96]. The authors find that sufficient contact tracing with quarantine can control a new outbreak of COVID-19. They mostly focus on the question of how much contacts need to be traced to control an epidemic for the three levels of basic reproductive number. However, this article neither explored the effectiveness of contact tracing for a specific location nor investigated the cost of contact tracing.

In this research, we develop an individual-based network framework to assess the impact of contact-tracing in the reopening process in a college town of Kansas. To analyze the cost of contact-tracing represented by the number of quarantined susceptible people, we develop a contact network and estimate the basic reproductive number R_0 and confirmed rate (infected to laboratory-confirmed transition) from observed confirmed case data in Manhattan, KS. We use our individual-based network model and the estimated parameters to run simulations of COVID-19 transmission. We use our framework to understand the COVID-19 propagation and assess the contact-tracing strategy in the different reopening situations.

Summarizing, the main contributions of this chapter are the following:

- A novel individual-level network-based epidemic model to assess the impact of contact tracing
- A thorough investigation of costs and benefits of contact-tracing in the reopening process in a college town of Kansas

The individual-based network model represents the heterogeneity in people mixing. Our individual-based network epidemic model is general and flexible to estimate and model contact-tracing for COVID-19 in any location. It also can model other diseases that have a similar transmission mechanism like COVID-19.

3.2 Data

The study area of this research is a college town in the rural region of the USA: Manhattan, KS. We use two data sets to develop our model. The first dataset contains the sociodemographic information from the census 2018, and the second dataset contains the COVID-19 incidence data. We also use Google COVID-19 community mobility reports [91] to reflect the "Stay-At-Home" situation.

3.3 Individual-based contact network model

We use demographic data to develop an individual-based contact network model capable of representing the heterogeneous social mixing. Our network has N nodes and L links. In this network, each node represents one occupied household, a link between two households represents the contact probability between members of these households. The system has a total population of p individuals, distributed randomly into the N occupied households according to five social characteristics: age, average household sizes, family households, couple, living-alone [90]. We maintain the average household sizes, number of family households, number of couples, and number of living-alone households. Besides, a person under 18 years old is always assigned to a house with at least one adult person. To develop this network, we consider five age-ranges: under 18, $18-24$, $25-34$, $35-59$, and over 60. Each age-range has p_i people, where $i \in \{1, 2, 3, 4, 5\}$. This model considers large shared living spaces (for example, dorms) as a set of households with 4-8 students in each household.

After assigning the people, an age-specific network is developed for each age range and a random mixing network for all ages. Then a combination of the six networks provides the full network. A full network represents a contact network for a typical situation. The configuration network model [97] is used to develop age-specific networks and the random mixing network. The system has N occupied households and p people. The steps to develop an age-specific network are:

Step 1: For each person j (here, $j \in 1, 2, ..., p$), contacts c_j is assigned from a Gaussian distribution $\mathcal{N}(\mu, \sigma^2)$. The mean μ of the Gaussian distributions are taken from the average number of daily contacts per person in each age-range [98–100]. The average daily contacts per person

are given in Table 3.1. For an under 18-year-old person, the number of contacts is assigned randomly from the $\mathcal{N}(13.91, 6.95)$ distribution. For a person in $18-24$ years age, the number of contacts is assigned randomly from the $\mathcal{N}(21.25, 10.62)$ distribution. For a person in $25-34$ years age, the number of contacts is assigned randomly from the $\mathcal{N}(21.3, 10.65)$ distribution. For a person in $35-59$ years age, the number of contacts is assigned from the $\mathcal{N}(20.912, 10.46)$ distribution. For an over 60-year-old person, the number of contacts is assigned randomly from the $\mathcal{N}(10.7, 5.35)$ distribution. In the random-mixing-network, the number of contacts is assigned randomly from the $\mathcal{N}(2, 1)$ distribution for a person j. The Gaussian or normal distribution is the distribution of real numbers; therefore, the number from the $\mathcal{N}(\mu, \sigma^2)$ distribution is rounded to the closest integer.

Step 2: For each person j, contacts for its belonging household k is assigned by $(c_j - h_k - 1)$. Here, c_j is the number of contacts for a person j, h_k is the household size or number of people of the household k, person j lives in the household k, $j = 1, 2, 3.....p$, and $k = 1, 2, 3.....N$.

Step 3: From the mixing patterns of different age-ranges, people have a strong tendency to meet people with their same age range (more than 80%) [98–100]. Therefore, We keep the maximum number of contacts among the same age ranges and a small percentage for the other age ranges. The percentage of contacts in the same age-specific-network for each age-range is given in Table 3.1. Degree d_{ki} of a node k in the age-specific network i is $s\%$ of $(c_j - h_k - 1)$, here, $s\%$ of average daily contacts of a person happens with the people of his same age-range.

Step 4: After assigning degree, d_{ki} for N nodes or households, The configuration network model [97] creates half-edges for each node, then chooses two nodes randomly and connect their half-edges to form a full edge [97].

The population and network characteristics for the five age-specific networks for Manhattan, KS are given in Table 3.1. According to census 2018, Manhattan, KS has $p = 55,489$ people and $N = 20,439$ occupied households [90].

Adjacency matrix for the full network A_f is a summation of six adjacency matrices: $A_f = \sum_{i=1}^{5} A_i + A_r$. Here, A_i is the adjacency matrix for the age-specific network i, and A_r is the adjacency ma-

Table 3.1: Properties of the Age-specific-networks of the Manhattan, KS.

Age-range	under 18	18-24	25-34	35-59	over 60
population	8074	20378	9887	10581	6567
average daily contacts per person [98]	13.91	21.25	21.3	20.91	10.7
average daily contacts with non-household members per person	12.00	20.00	19.98	19.00	7.05
% of neighbors in the same age-specific networks [99]	85.63	90.48	90.29	84.95	71.43
number of edges in the age-specific networks	40466	187723	88806	90835	16511

trix for the random mixing network. Age-specific networks and the random mixing network are unweighted and undirected. However, the full network is weighted and undirected. The full network for Manhattan (KS) has $445,350$ edges. The average node degree for an individual household in the full-network is 43.647, and for an individual person is 16.0518 (which is consistent with [98]). The degree distribution is presented in Fig. 3.1. The networks are available at https://doi.org/10.7910/DVN/3IM82E.

The full network is a contact network in the normal situation; we modify it to represent the contact network in the pandemic lockdown; we name it *limited network*. Manhattan, KS, is the home of Kansas State University. Most of the people living in Manhattan, KS, are closely related to Kansas State University, which halted its in-person activities from early March 2020 to August 17, 2020. Besides, Manhattan, KS was under the "Stay-At-Home" order from March 27, 2020, to May 4, 2020 [101]. To represent this unusual situation, we modify the full network to a limited network version. As the educational institute was closed, we randomly reduce 90% links from the age-specific networks for the age-ranges under 18 and $18-24$. The Google COVID-19 community mobility reports provide a percentage of movement changes in different places

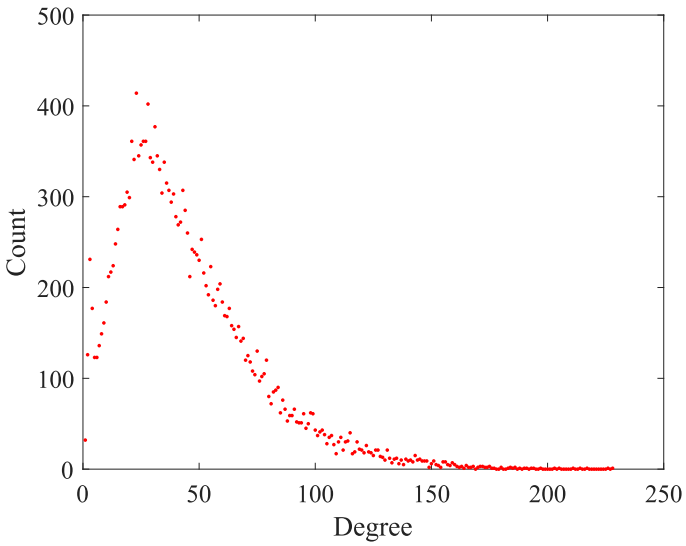


Figure 3.1: **Degree distribution of the full network.** In the network, households are at the node level. The network has 20,439 nodes and 445,350 edges. The average degree of this network is 43.647. The maximum degree in the network is 227.

(for example, workplaces, recreational areas, parks) [91]. We reduced 40% links randomly from the age-specific networks for 25 – 34, and 35 – 59 age-ranges for the movement changes in the workplaces [91]. The number of links in the limited network is 155,762. The limited network is available at https://doi.org/10.7910/DVN/3IM82E.

3.4 Epidemic model

We design a susceptible-exposed-infected-confirmed-removed (SEICR) epidemic scheme to simulate the COVID-19 transmission (Fig. 3.2). This model has five compartments: susceptible S, exposed E, infected I, confirmed C, and removed R. A susceptible node is a node that is not infected yet. An exposed node is a node infected by the disease, but the viremia level is deficient that it cannot infect other nodes. An infected node is infectious, and it can infect other nodes. In this model, an infected node can be symptomatic, asymptomatic, or presymptomatic. A confirmed node is a laboratory-confirmed COVID-19 case. A removed node can be recovered or dead. The SEICR model has five transitions, which are divided into two categories: edge-based ($S \rightarrow E$),

and nodal ($E \rightarrow I$; $I \rightarrow C$; $C \rightarrow R$; $I \rightarrow R$) transitions [1, 19].

An edge-based transition of a node depends on the state of its contacting nodes or neighbors in the contact network with its own state. A nodal transition of a node only depends on the own state. Each edge-based transition has an influencer compartment. A transition from susceptible to exposed ($S \rightarrow E$) of a susceptible node depends on the infected neighbors of that node. Therefore it is an edge-based transition, and the infected compartment is the influencer compartment of this transition. In this work, we are using the term 'neighbors of a node k' for the nodes, which have the shortest path length 1 from the node k. The transition rate of the susceptible to exposed ($S \rightarrow E$) transition of a node k is $\beta_1 \sum_{l}^{N} A_c(k,l) I_l$, here, β_1 is the transmission rate from one infected node to one susceptible node, A_c is the adjacency matrix of the contact network, if l node is infected then $I_l = 1$ otherwise $I_l = 0$, and $\sum_{l}^{N} A_c(k,l) I_l$ is the number of infected neighbors of the node k. The transition rate for the transition exposed to infected ($E \rightarrow I$) is δ_1. The confirmed rate of an infected person is δ_2. We consider that a laboratory-confirmed case will be isolated and cannot transmit the disease outside of his household anymore. The unknown COVID-19 cases will move from infected to removed with a rate $\delta_2^{'}$. We add another transition $C \rightarrow R$ with rate δ_1, this transition does not have any significance in the disease transmission. All the transition rates are exponentially distributed with a constant average value (Table 4.2). A detail of the SEICR epidemic model is stated in Table 4.2.

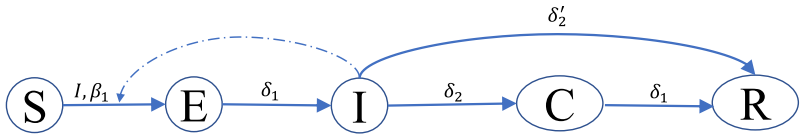


Figure 3.2: **Node transition diagram of the susceptible-exposed-infected-confirmed (SEICR) epidemic model.** This model has five compartments: susceptible (S), exposed (E), infected (I), confirmed (C), and removed (R) compartments. The SEICR model has five transitions (presented by solid lines): $S \rightarrow E$ (edge-based), $E \rightarrow I$ (nodal), $I \rightarrow C$ (nodal), $C \rightarrow R$ (nodal), and $I \rightarrow R$ (nodal). The infected (I) compartment is the influencer compartment of the edge-based $S \rightarrow E$ transition. The dashed line presents the influence of the I compartment on the $S \rightarrow E$ transition. We estimate R_0 and δ_2 transition rate from data. We deduce β_1 from R_0.

Table 3.2: Description of the susceptible-exposed-infected-confirmed (SEICR) epidemic model.

States	type	transition	average transition rate ($days^{-1}$)	influencer	source
S (Susceptible) E (Exposed) I (Infected) C (Confirmed)	Edge-based	$S \rightarrow E$	$\beta_1 \sum_{l}^{N} A_c(k,l)I_l$ here, $\beta_1 = \frac{R_0\delta_2}{\langle d\rangle\langle w\rangle}$; $\langle d\rangle$ = average degree; $\langle w\rangle$ = average weight	Neighbors in state I	R_0 is estimated
	nodal	$E \rightarrow I$	$\delta_1 = \frac{1}{3}$	-	[102, 103]
		$I \rightarrow C$	$\delta_2 = \frac{1}{4.56}$	-	estimated
		$C \rightarrow R$	$\delta_1 = \frac{1}{3}$	-	model
		$I \rightarrow R$	$\delta_2^{'} = 0.66\delta_2$	-	[104]

3.4.1 Parameter estimation for the SEICR epidemic model

The SEICR model has two unknown parameters: basic reproductive number R_0, and confirmed or reporting rate δ_2. To estimate the R_0 and δ_2, we have used confirmed cases in Riley County (Kansas) from March 25, 2020 to May 4, 2020. In this period, Kansas State University was closed, and "Stay-At-Home" order was there. It is reasonable to use this time period to estimate R_0 as there was no reopening and the mobility was the same throughout the period in Manhattan, KS. For the simulation of this period, a limited network is used, which is a modified version of the Full network to simulate the particular situation under the "Stay-At-Home" order. We use approximate Bayesian computation based on sequential Monte Carlo sampling (ABS-SMC) approach to estimate R_0 and δ_2 [11, 30]. The algorithm is in Appendix A. Other parameters (δ_1 [102, 103], and $\delta_2^{'}$ [104]) are taken from the literature.

The estimated value for R_0 is 0.55 (95% confidence interval: $0.522 - 0.564$) and for reporting rate δ_2 is $\frac{1}{4.79}\text{day}^{-1}$ (95% confidence interval: $\frac{1}{4.89} - \frac{1}{4.74}\text{day}^{-1}$). These estimated values are specific for Manhattan, KS for the time from March 25, 2020 to May 4, 2020. The R_0 for different reopening scenarios is presented in the supplementary Fig. S1. We consider that some people will develop severe symptoms, and they will be reported as a confirmed case of COVID-19 sooner. However, some people will produce deficient symptoms, and may they will be tested later. Therefore, the estimated confirmed rate is an average of all possibilities.

A sensitivity analysis for R_0 and reporting time on the mean-squared error between confirmed cases

data and simulated results is presented in Fig. 3.3.

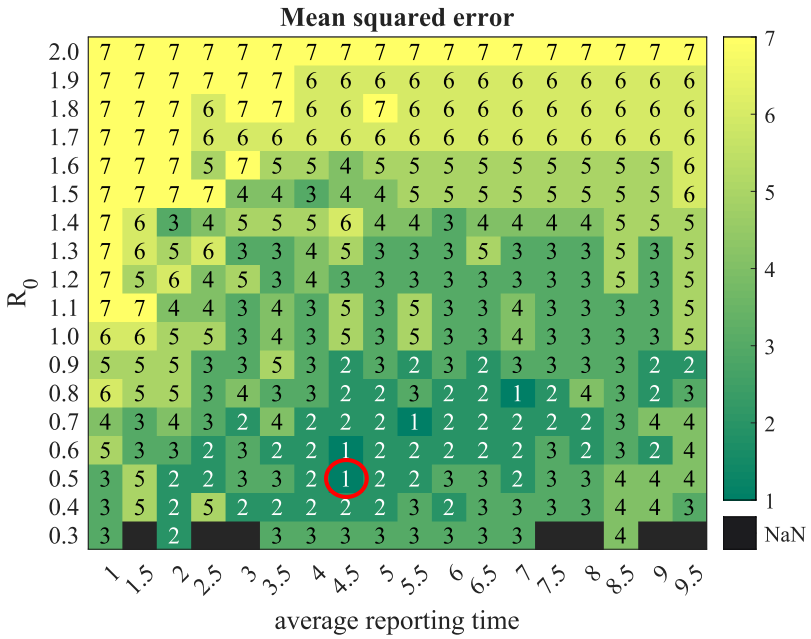


Figure 3.3: **A sensitivity analysis.** Mean-squared error (mse) between the time series of the total confirmed cases (or cumulative new cases per day) of March 25, 2020 to May 4, 2020 and simulated results for a different combination of basic reproductive number and average reporting time (in days). The light-colored boxes represent more mse than dark-colored boxes. The color boxes with number "1" means that mse$\leq$ 3, number "2" means that 3 <mse $\leq$ 10, number "3" means that 10 <mse $\leq$ 50, number "4" means that 50 <mse $\leq$ 100,number "5" means that 100 <mse $\leq$ 500,number "6" means that 500 <mse $\leq$ 1000,number "7" means that 1000 <mse. More than 80% times epidemic dies out in the combinations of the black squares, and confirmed cases are less than 10. The minimum error combination is showing by the red circle. We estimate R_0 = 0.55 and average reporting time= 4.79 days.

3.4.2 Simulation for four different reopening scenarios

We simulate the total confirmed cases (or cumulative new cases per day) for eight months: from May to December using the SEICR epidemic model with the estimated parameters. To simulate, we assume that there is no change except reopening from pandemic lockdown. We are presenting four reopening situations: "Stay-At-Home" is still there or no reopening, 25% reopening, 50% reopening, and 75% reopening. Kansas has started to reopen step by step after May 4, 2020. We use the limited network to simulate from March 25, 2020 to May 4, 2020; then, we change the network concerning the reopening situation. For example, in a 25% reopening situation, 25% of

the reduced movement will start again; to model it, we add 25% missing links randomly (which are present in the full network but not in the limited network). We preserve the states of each node at May 4, 2020 in the network then use it as the initial condition for the simulation for the reopening situation (from May 4, 2020 to July 1, 2020). Fig. 3.4 is showing the medians (solid lines) and interquartile ranges (shaded regions) of the total confirmed cases of the 1000 stochastic realizations of the four reopening scenarios. The zoom-in window in Fig. 3.4 shows the time period when data was used to estimate the parameters of the epidemic model.

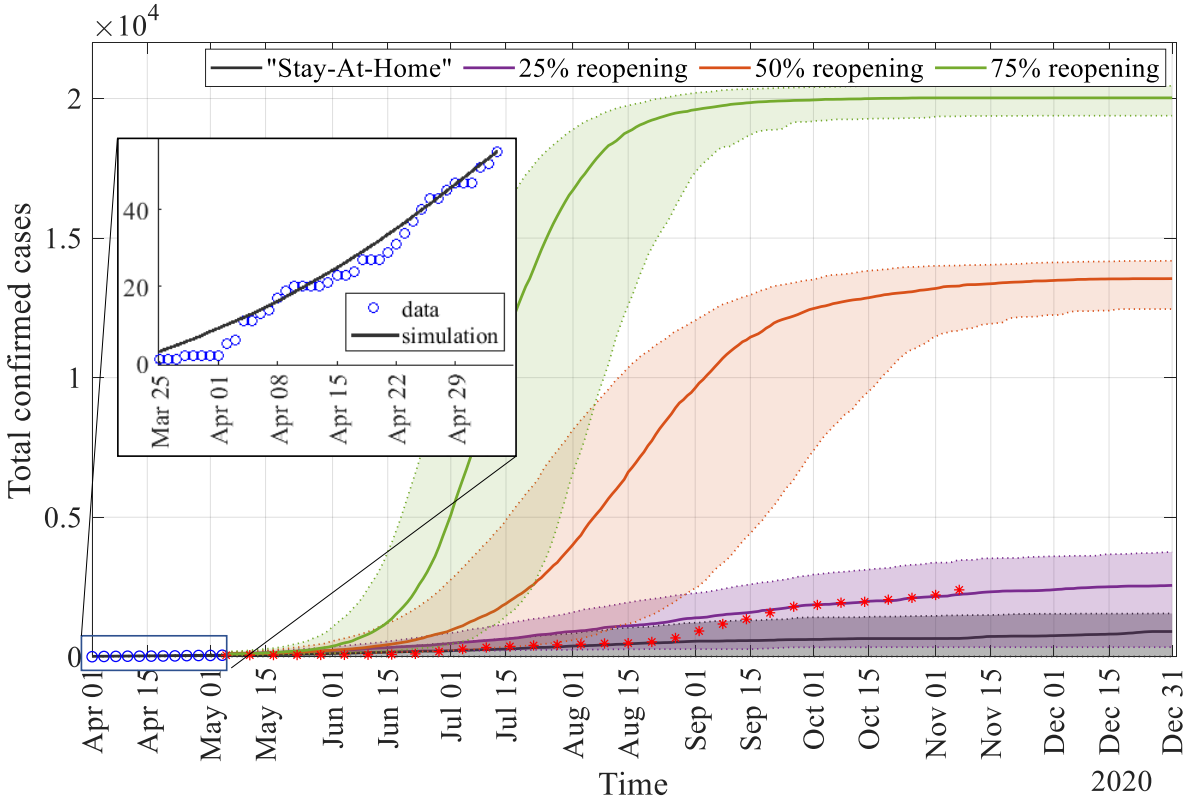


Figure 3.4: **Total confirmed cases with time in the four reopening scenarios after 'stay at home' order lifted on May 4, 2020.** Solid lines represent the median, and shaded regions represent interquartile range of the 1000 stochastic realizations. The blue circles in the zoom-in window present the total confirmed case data of the COVID-19 in Manhattan (Kansas) for the time period from March 25, 2020 to May 4, 2020. We have used this time period to estimate the basic reproductive number and the average confirmed time. The red stars are the total confirmed case data of the COVID-19 in Manhattan (Kansas) after May 5, 2020.

3.4.3 Stochastic simulation

To do the simulation, we use GEMFsim; it is a stochastic simulator for the generalized epidemic modeling framework (GEMF), which was developed by the Network Science and Engineering (NetSE) group at Kansas State University [105]. The GEMFsim is a continuous-time, individual-based, numerical simulator for the GEMF-based processes [1]. The network and epidemic model is the input of the GEMFsim, and the time dynamic of each node state is the output. In GEMF, the joint state of all nodes follows a Markov process that arises from node-level transition. A node can change its state by moving from one compartment to another compartment through a transition. One assumption of the GEMF system is, all the events or transitions are independent Poisson processes with a constant rate; this assumption leads the system to a continuous-time Markov process. Initially, the simulation starts by setting two infected nodes randomly. The stochastic simulator GEMFsim is based on the Gillespie algorithm. The Gillespie algorithm can produce a statistically correct trajectory of a continuous-time Markov process.

3.5 Contact tracing

Contact tracing is a key mitigation strategy to control the COVID-19 propagation. To implement contact tracing, we modify the basic SEICR epidemic model and propose a two-layer network model. In the implementation of the contact tracing, we follow the CDC's guidance for contact tracing [95].

3.5.1 Two-layer individual-based network model

This work implements contact tracing in a two-layer network model: the contact network is in the first layer, and the tracing network is in the second layer (Fig. 3.5). We will call the first layer as the contact-layer and second layer as the tracing-layer in the rest of the chapter. In the t%-tracing-layer, t% of links of each node in the contact-layer are preserved randomly. To form a t%-tracing-layer, at first, we generate a random number r from $U(0,1)$ for each link from a node i; then keep the link in the tracing-layer if $r \leq 0.01t$. A 50% tracing-layer is presented in Fig.

3.5. Although the contact-layer is an undirected network, however, the tracing-layer is a directed network. In the directed tracing-layer, a neighboring node of a node i has a distance one from node i. The neighbors of a confirmed (C) node in the tracing-layer will be tested and quarantined.

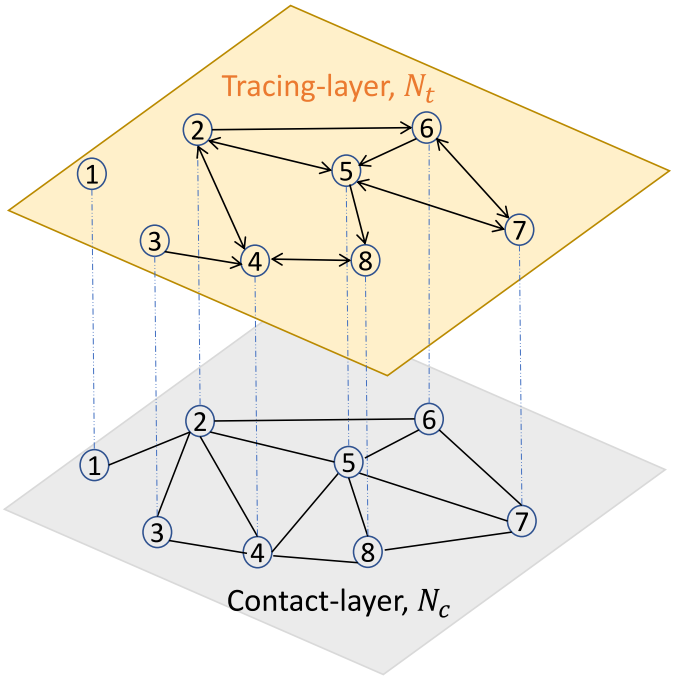


Figure 3.5: **Two-layer network model: contact-layer N_C, and tracing-layer N_t.** In this example, 50% of contacts of each node is traced. For example, node 4 has four neighbors in the contact-layer (2, 3, 5, 8); however, two neighbors in the tracing-layer (2, 8). Node 7 has three neighbors in the contact-layer (6, 5, 8); however, two neighbors in the tracing-layer (6, 5). Node 8 has three neighbors in the contact-layer (4, 5, 7); however, one neighbor in the tracing-layer (4).

3.5.2 Epidemic scheme for contact tracing

For the contact tracing mitigation strategy, we consider two approaches for quarantine: I) all the neighbors of a confirmed case in the tracing-layer will be quarantined, and II) only infected neighbors of a confirmed case in the tracing-layer will be isolated. For the case I, we propose the SEICQ1 epidemic model, and for case II, we propose the SEICQ2 epidemic model. The SEICQ1 model has eight compartments: susceptible (S), exposed (E), infected (I), confirmed (C), quarantined-susceptible (Q_S), quarantined-exposed (Q_E), quarantined-infected (Q_I), and removed (R). The SEICQ2 model has six compartments: susceptible (S), exposed (E), infected (I), confirmed (C),

quarantined-infected (Q_I), and removed (R). The transitions $S \rightarrow E$, $E \rightarrow I$, $I \rightarrow C$, and $I \rightarrow R$ are the same as the base SEICR model.

In the SEICQ1 model, neighbors (susceptible, exposed, and infected) of a confirmed node in

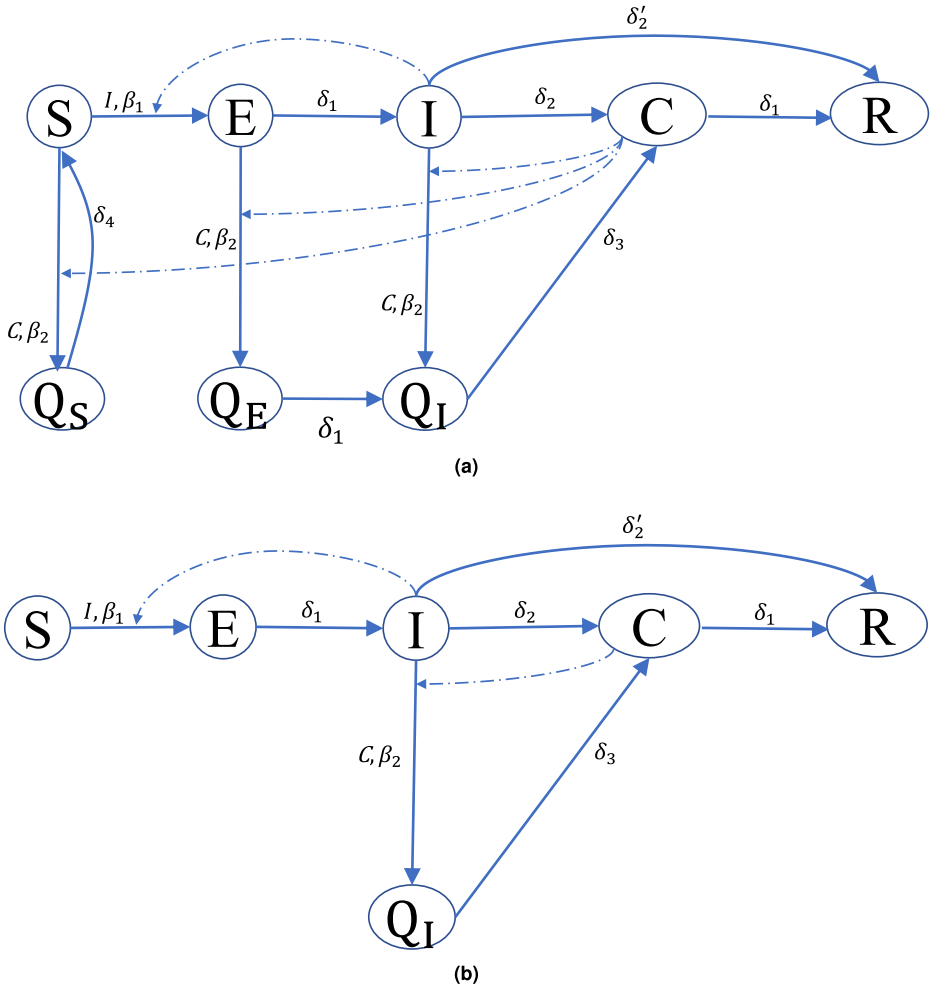


Figure 3.6: **Node transition diagrams.** a) SEICQ1 epidemic model, b) SEICQ2 epidemic model. The solid lines represent the node-level transitions, and the dashed lines represent the influence of the influencer compartment on an edge-based transition.

the tracing-layer will be tested and quarantined. In the SEICQ1 model, susceptible, exposed, infected neighbors in the tracing-layer of a confirmed case will go to the quarantined-susceptible Q_S, quarantined-exposed Q_E, and quarantined-infected Q_I states with rate β_2. The susceptible to quarantined-susceptible ($S \rightarrow Q_S$), exposed to quarantined-exposed ($E \rightarrow Q_E$), and infected to quarantined-infected ($I \rightarrow Q_I$) transitions are edge-based transitions and confirmed compartment is the influencer of these transitions. A COVID-19 positive neighbor of a confirmed node will

go to the confirmed state immediately with δ_3 rate, $Q_I \rightarrow C$ is a nodal transition. We model the transition rates β_2 and δ_3 are much higher than the rate for the transition $C \rightarrow R$ to ensure that the neighbors of a confirmed node in the tracing-layer will move to the quarantined or confirmed state before the $C \rightarrow R$ event happens. For the simulation, we take $\beta_2 = \delta_3 = 50\delta_1$. The SEICQ1 model is presented in Fig. 3.6a. A description of the 11 transitions of the SEICQ1 model is given in Table 3.3.

In the SEICQ2 model, neighbors of a confirmed node in the tracing-layer will be tested, and only infected neighbors will go to the quarantined-infected (Q_I) state immediately with the rate β_2. The node transition diagram of the SEICQ2 model is given in Fig. 3.6b. A description of the seven transitions of the SEICQ2 model is given in Table 3.4.

Table 3.3: Description of the SEICQ1 epidemic model.

States	type	transition	transition rate ($days^{-1}$)	inducer	source
S (Susceptible) E (Exposed) I (Infected) C (Confirmed) Q_S (Quarantined-Susceptible) Q_E (Quarantined-Exposed) Q_I (Quarantined-Infected) R (Removed)	edge-based	$S \rightarrow E$	$\beta_1 \sum_{l}^{N} A_c(k,l) I_l$ here, $\beta_1 = \frac{R_0 \delta_2}{\langle d \rangle \langle w \rangle}$; $\langle d \rangle$ = average degree; $\langle w \rangle$ = average weight	Neighbors of state I in the contact-layer	R_0 is estimated
		$S \rightarrow Q_S$ $E \rightarrow Q_E$ $I \rightarrow Q_I$	$\beta_2 \sum_{l} A_t(k,l) C_l$ here, $\beta_2 >> \delta_1$, we take $\beta_2 = 50\delta_1$	Neighbors of state C in the tracing-layer	model
	nodal	$E \rightarrow I$ $Q_E \rightarrow Q_I$	$\delta_1 = \frac{1}{3}$	-	[102, 103]
		$C \rightarrow R$	$\delta_1 = \frac{1}{3}$	-	model
		$I \rightarrow C$	$\delta_2 = \frac{1}{4.56}$	-	estimated
		$I \rightarrow R$	$\delta_2' = 0.66\delta_2$	-	[104]
		$Q_I \rightarrow C$	$\delta_3 >> \delta_1$, we take $\delta_3 = 50\delta_1$	-	model
		$Q_S \rightarrow S$	$\delta_4 = \frac{1}{14}$	-	[95]

.

Table 3.4: Description of the SEICQ2 epidemic model.

States	type	transition	transition rate ($days^{-1}$)	inducer	source
S (Susceptible) E (Exposed) I (Infected) C (Confirmed) Q_I(Quarantined-Infected) R (Removed)	edge-based	$S \rightarrow E$	$\beta_1 \sum_{l}^{N} A_c(k,l)I_l$ here, $\beta_1 = \frac{R_0\delta_2}{\langle d\rangle\langle w\rangle}$; $\langle d\rangle$ = average degree; $\langle w\rangle$ = average weight	Neighbors of state I in the contact-layer	R_0 is estimated
		$I \rightarrow Q_I$	$\beta_2 \sum_{l} A_t(k,l)C_l$ here, $\beta_2 >> \delta_1$, we take $\beta_2 = 50\delta_1$	Neighbors of state C in the tracing-layer	model
	nodal	$E \rightarrow I$	$\delta_1 = \frac{1}{3}$		[102, 103]
		$C \rightarrow R$	$\delta_1 = \frac{1}{3}$	-	model
		$I \rightarrow C$	$\delta_2 = \frac{1}{4.56}$	-	estimated
		$I \rightarrow R$	$\delta_2^{'} = 0.66\delta_2$	-	[104]
		$Q_I \rightarrow C$	$\delta_3 >> \delta_1$, we take $\delta_3 = 50\delta_1$	-	model

3.5.3 Impact of contact tracing

Contact tracing can minimize the effect of the reopening process and control the COVID-19 transmission. We apply contact tracing after May 4, 2020 in Manhattan, KS. The simulation plot of total confirmed cases on December 31, 2020 is presented in Fig. 3.7 for four reopening scenarios : 25% reopening, 50% reopening, 75% reopening, and 100% reopening for the different levels of contact tracing. The solid lines in Fig. 3.7 represent the median, and shaded regions represent the interquartile range of the 1000 stochastic realizations for the SEICQ1 and SEICQ2 model.

The difference between SEICQ1 and SEICQ2 is that SEICQ1 quarantines susceptible, exposed, and infected neighbors of a confirmed case in the tracing-layer; however, SEICQ2 isolates only the infected neighbors of a confirmed case in the tracing-layer. The SEICQ1 model is always more efficient than the SEICQ2 model to control the COVID-19 propagation. However, both approaches can reduce the number of confirmed cases, even in the 100% reopening situation. For any reopen-

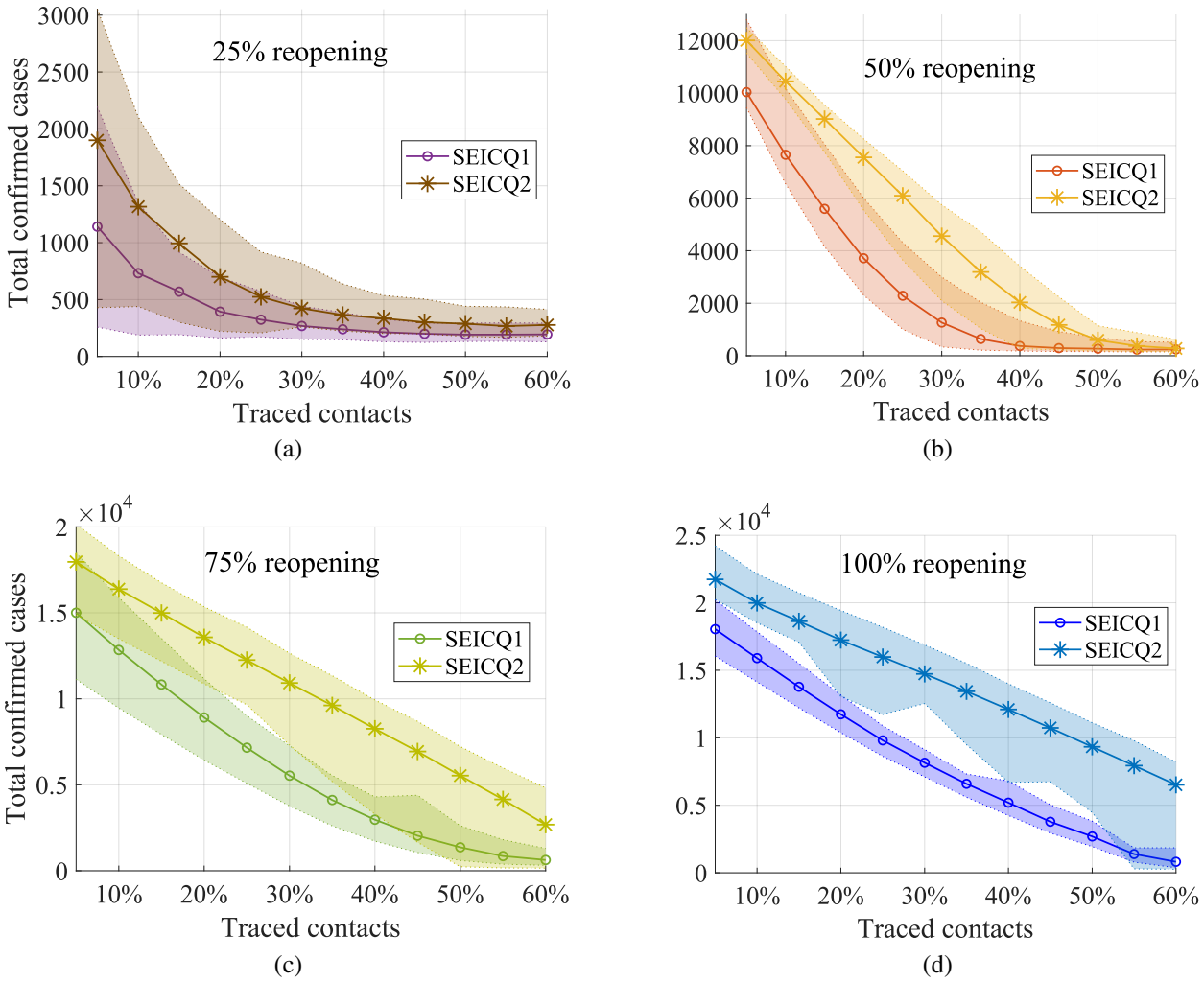


Figure 3.7: **Impact of contact tracing.** Total reported cases in eight months after 'Stay-At-Home order' lifted for different movement restrictions scenarios. Contact tracing is applied after May 4, 2020. This figure is showing the median (solid lines) and interquartile range (shaded regions) value of 1000 stochastic realizations.

ing situations, tracing more than 55% of the contacts in the SEICQ1 can reduce the median of the 1000 stochastic realizations of the confirmed cases more than 90%, and in the SEICQ2 can reduce the median of the 1000 stochastic realizations of the confirmed cases more than 66% on December 31, 2020, with compare to no-contact-tracing (SEICR model) (Table 3.5).

The SEICQ1 model can reduce the reported cases further compared to SEICQ2 for the same amount of contact tracing (Fig. 3.7). However, the SEICQ1 model has a drawback; it quarantines susceptible persons. The number of total quarantined susceptible households in the simulation time period for different amounts of traced contacts for the SEICQ1 model is presented in Fig. 3.8 and

Table 3.5: Percentage of reduction of the total confirmed cases in eight months after May 4, 2020, in the four reopening scenarios for the two contact tracing mitigation approaches.

Traced contacts	Percentage of reduction in the total confirmed cases							
	SEICQ1				SEICQ2			
	25% re-opening	50% re-opening	75% re-opening	100% reopening	25% re-opening	50% re-opening	75% re-opening	100% reopening
5%	55.38	25.90	25.06	24.52	25.70	11.36	10.28	9.10
10%	71.37	43.50	35.87	33.5	48.54	22.84	18.20	16.40
15%	77.73	58.7	45.91	42.39	61.14	33.44	25.14	22.08
20%	84.60	72.60	55.49	42.39	72.64	44.24	32.21	27.90
25%	87.35	83.14	64.26	58.98	79.51	55.01	38.80	33.14
30%	89.52	90.67	72.37	65.90	83.47	66.36	45.47	38.39
35%	90.65	95.29	79.45	72.43	85.74	76.46	51.99	43.80
40%	91.67	97.26	85.16	78.28	86.90	85.00	58.76	49.40
45%	92.20	97.86	89.82	84.19	88.23	91.34	65.36	55.11
50%	92.49	98.02	93.18	88.74	88.75	95.62	72.36	60.97
55%	92.48	98.27	95.73	94.19	89.54	97.25	79.28	66.80
60%	92.41	98.22	96.87	96.60	89.17	97.96	86.59	72.75

Table 3.6. The quarantined susceptible households increase with the increase of tracing; however, after tracing a certain percentage (t_p%) of contacts, the quarantined susceptible households start to decrease with the increase of tracing (Fig. 3.8). If we consider quarantined susceptible households are the cost of SEICQ1 model, then it is cost-effective to trace contacts of the confirmed cases more than t_p%; which is 10% for 25% reopening, 10% for 50% reopening, 20% for 75% reopening, and 25% for 100% reopening (Table 3.6). The reason for decreasing the number of quarantined households with the increasing of contact-tracing after the t_p% is the smaller number of infected cases. Although each confirmed case will give a long list of possible contacts, this effect will be balanced out by a decreasing number of the confirmed cases (supplementary Fig. S2-S5).

3.6 Summary

This research studies contact tracing as a key mitigation strategy to control the COVID-19 transmission in the reopening process of a college town in the rural region of the USA. Therefore, we

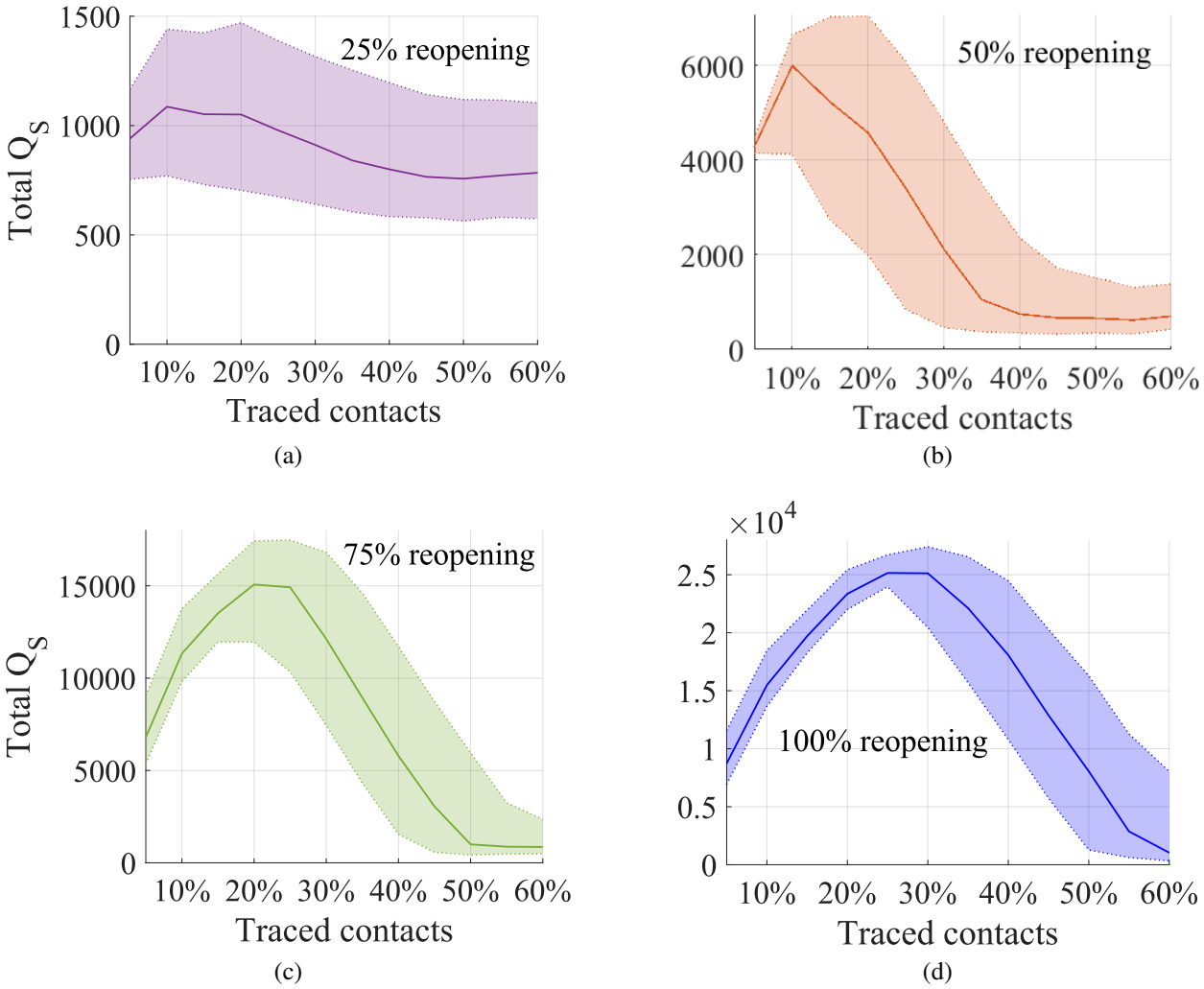


Figure 3.8: **The total number of quarantined susceptible households** in eight months after May 4, 2020, for the SEICQ1 epidemic model for the four reopening scenarios with different tracing levels. This figure is showing the median (solid lines) and interquartile range (shaded regions) of 1000 stochastic realizations.

Table 3.6: Total quarantined susceptible households in eight months after May 4, 2020, in the SEICQ1 epidemic model for the four reopening scenarios.

Traced contacts	SEICQ1							
	25% reopening		50% reopening		75% reopening		100% reopening	
	Total confirmed cases (median)	Total quarantined Susceptible households (median)	Total confirmed cases (median)	Total quarantined Susceptible households (median)	Total confirmed cases (median)	Total quarantined Susceptible households (median)	Total confirmed cases (median)	Total quarantined Susceptible households (median)
5%	1141	941	10039	4278	15005	6796	18048	8727
10%	732	1086	7653	5997	12840	11329	15899	15486
15%	569	1052	5594	5231	10829	13515	13775	19706
20%	393	1051	3712	4582	8911	15071	11740	23355
25%	323	978	2284	3401	7156	14919	9807	25150
30%	268	911	1263	2124	5531	12120	8151	25122
35%	239	840	638	1053	4113	8938	6589	22120
40%	213	800	370	748	2970	5791	5193	18069
45%	199	765	290	667	2039	3050	3780	12883
50%	192	757	267	659	1365	998	2693	8052
55%	192	772	233	622	853	872	1386	2870
60%	194	784	241	650	625	856	813	1037

propose a general framework to develop an individual-based contact network epidemic model to estimate parameters and implement contact tracing. This model is used to estimate the basic reproductive number (R_0) and confirmed rate (δ_2) in Manhattan, KS, for the COVID-19 propagation. The outcomes of this research are valuable to understand the effectiveness of the contact-tracing strategy in the different scenarios of the COVID-19 transmission. Furthermore, this framework is generic enough to use any locations and for other diseases as well.

The individual-based network model represents the heterogeneous mixing nature of a population. To investigate transmission at the individual level, we develop an individual-based contact network model where households are presented by network nodes. The contact network is a combination of five age-specific networks and one random-mixing network; this approach allows us to change an age-specific network according to any change in the society (for example, summer break, pandemic lockdown). The pandemic lockdown reduces the contacts mostly among the people who are students. Therefore, age-specific networks for under 18 and 18-24 are changed mostly. Pandemic lockdown also affects people in 25-34, 35-59 age-ranges. We propose a 'full network' to represent the usual situation; then, we modify the age-specific networks of the full network according to Google COVID-19 community mobility reports [91] to represent pandemic lockdown . The modified network is the limited network, a reduced version of the full network. The average degree of the full network is 43.647 for Manhattan, KS which means that each household has probable direct connections with an average of 43.647 households. The full network is connected and provides an approximation of the contact network at the household level, which is useful for doing the simulation anonymously.

We propose a susceptible-exposed-infected-confirmed-removed (SEICR) epidemic model in the limited network to simulate COVID-19 transmission from March 25, 2020 to May 4, 2020. We estimate the unknown parameters of the SEICR model for the Manhattan, KS, using approximate Bayesian computation based on sequential Monte Carlo sampling. We use confirmed cases as an observed data set. Designing an optimal epidemic model to simulate epidemic processes is essential. However, it is challenging to design an epidemic model for COVID-19 transmission with limited knowledge; understanding the COVID-19 transmission needs more investigation. Concerning the unclear role of immunity, we assume that the immunity of a recovered COVID-19

patient is not going to fade in the short period analyzed in our simulations. In addition, we assume that a tested positive person is responsible enough to stay in isolation. However, it is important to keep the model simple, since the data available to estimate parameters is limited. Therefore, we propose a simple but dynamic and flexible epidemic model to simulate COVID-19 transmission, which has only two unknown parameters. The model can easily cope with additional information that may be available in the future.

The estimated basic reproductive number is much smaller in Manhattan, KS (estimated $R_0 = 0.55$) because of the 'Stay at home' order. In Manhattan, 51% of people have age below 24 years, who get a chance to stay at home because of the online curriculum in educational institutions. However, the basic reproductive number will change when educational institutes start their in-person curriculum (in the 100% reopening, the deduced R_0 is 2.0301). There are 301 college towns in the USA [106], which have a similar population structure like Manhattan, KS. A practical contact tracing approach can help to control the epidemic in those college towns.

We implement contact tracing using a two-layer network model. We assess the impact of contact tracing in the four reopening scenarios: 25% reopening, 50% reopening, 75% reopening, and 100% reopening (or no restrictions). Reopening without vaccination is challenging. It is essential to access the efficacy of the contact tracing in the reopening path. Our investigation indicates that more than 50% contact tracing can control the COVID-19 transmission even in the 100% reopening situation. The number of quarantined susceptible people increases with the increase of traced contacts, however after a certain amount of tracing (t_p%), the number of quarantined susceptible people decreases with the increases of the traced contacts. We consider that quarantined susceptible people represent the cost of SEICQ1 contact tracing model. Therefore, it is cost-effective to trace more than t_p% contacts of a confirmed case. Our research finds that t_p increases with the increase in mobility (Table 3).

Our investigation indicates that a sufficient amount of contact tracing can reduce the COVID-19 transmission in the reopening process of a location. At first, the quarantined susceptible people increase with the percentage of traced contacts, however after a certain amount of traced contacts, the quarantined susceptible people start to decrease with the increase in the percentage of traced contacts.

3.7 Data availability

The dataset and code used to perform this research is available from https://doi.org/10.7910/DVN/3IM82E. The authors are willing to provide further details upon request.

Chapter 4

Group-based general epidemic modeling for spreading processes on networks: GroupGEM[1]

We develop a group-based continuous-time Markov general epidemic modeling (GroupGEM) framework for any compartmental epidemic model (e.g., susceptible-infected-susceptible, susceptible-infected-recovered, susceptible-exposed-infected-recovered). Here, a group consists of a collection of individual nodes of a network. This model can be used to understand the critical dynamic characteristics of a stochastic epidemic spreading over large complex networks while being informative about the state of groups. Aggregating nodes by groups, the state-space becomes smaller than the one of individual-based approach at the cost of an aggregation error, which is bounded by the well-known isoperimetric inequality. We also develop a mean-field approximation of this framework to reduce the state-space size further. Finally, we extend the GroupGEM to multilayer networks. Individual-based frameworks are in general not computationally efficient. However, the individual-based approach is essential when the objective is to study the local dynamics at the individual level. Therefore, we propose a group-based framework to reduce the computational time of the Individual-based generalized epidemic modeling framework (GEMF) but retain its advantages.

[1]This chapter is a slightly modified version of our published article [19], Copyright ©2021, IEEE Transactions on Network Science and Engineering.

4.1 Background

Epidemic spreading processes over complex networks is an essential topic for different research fields such as epidemiology, social science, and computer science [6, 24, 69, 107]. Theoretical models of stochastic epidemic spreading processes over a network can reveal important dynamic characteristics of an epidemic. The spread of computer viruses, information, opinions, rumors, knowledge, products, or any spreading process in a network of interactive agents can be modeled as an epidemic process. All the above spreading processes follow some common patterns.

Compartmental models are widely used in the study of epidemics. In a compartmental model, individuals/agents can be in different compartments. The set of compartments can be different for different models. Widely used compartments in the literature are susceptible, infected, recovered, immune, and latent [69]. The compartments can be different for different research areas or scenarios. An individual can move from one compartment to another. In this research, we assume that this event is an independent Poisson process with a constant rate; this assumption leads the system to a continuous-time Markov process.

Some complex networks have a large set of nodes/agents, and epidemic modeling over those very large networks is computationally expensive and time-consuming. To address this issue, researchers have proposed several impactful models; Volz's probability generating function (PGF) model [108], Miller's edge-based compartmental (EBC) model [109], and Lindquist's effective degree model [110]. Those models are computationally efficient. However, those models can only provide information about the aggregate dynamics of an epidemic. In addition, most of those models are specific to a certain epidemic model on a specific category of networks [108–111]. The individual-based approach can allow us to understand the local dynamics of an epidemic. The individual-based approach is also more flexible in terms of the initial conditions. Sahneh et al. proposed a generalized epidemic modeling framework (GEMF) for the individual-based approach [1]. However, a drawback of the individual-based approach is computational time. To reduce the computational time of the individual-based approach, sometimes researchers scale their population by considering several individuals or a group as a single individual node [11, 36, 37]. This type of scaling can alter the actual system, and estimation of the dynamics can be misleading. In this

chapter, we propose a group-based model to reduce the computational time of the individual-based framework (GEMF) while retaining its advantages.

Grouping or partitioning of the nodes of a network can be user-defined, deterministic from data, or random. A group-state can tell us the summary of its nodes states and provide the local dynamics at the group-level. A group-based framework is useful to find out the dynamics of any static network such as communication, trade, social, biological, livestock, or power-grid networks. In this new era, because of the improvement of digital technology, different types of communication among humans are popular. As a result, a vast number of people are connected to form very large networks. These networks can influence public opinion, which is very impactful in the field of politics, economy, business, and others. A group-based framework can be useful to understand the different dynamics of these very large networks. It can also help us find out the impact of a group on the dynamics of the system. For example, it can assist in understanding the influence of the opinion of a social group in a political event such as an election. The group-based approach can be useful as well in modeling co-evolution spreading. In this big data age, it is possible to develop realistic data-driven co-evolution spreading networks [112], and it is crucial to have tools to handle them.

The heterogeneous mean-field method (HMF) [113, 114] and the N-intertwined mean-field method (NIMFA) [14, 115] (also called the quenched mean-field (QMF), or individual-based mean-field (IBMF) method [6]) are two well-established approximation methods for analysis of dynamical processes on complex heterogeneous networks. They are two particular cases of the group-based unified mean-field framework (UMFF), which was first proposed by Devriendt et al. for the susceptible-infected-susceptible epidemic model [28]. In this article, we generalize the model of Devriendt et al. and develop a group-based continuous-time Markov epidemic modeling (GroupGEM) framework. The group-based approach has fewer degrees of freedom than NIMFA. Although HMF also has this property, UMFF has more flexibility to choose groups. The heterogeneous mean-field method (HMF) is a degree-based approach, and nodes of the same degree are assumed statistically equivalent, which is not the only case for UMFF. Several models were developed to improve HMF and IBMF approximations. The dynamical message-passing (DMP) model and pair-quenched mean-field model were developed to improve the individual-based mean-field ap-

proximation. However, HMF, IBMF, and DMP models cannot capture network topology and dynamical correlations together [116]. The pair quenched-mean-field (pQMF) model considers the dynamical correlation between connected nodes by using pair approximation, and its equations can be obtained by higher-order group-based UMFF equations, presented in detail in ref. [28]. The accuracy of all mean-field models is affected by a moment-closure approximation. We develop a group-based continuous-time Markov epidemic modeling (GroupGEM) framework, which does not contain a moment-closure approximation. Then, we propose mean-field equations of this framework (GroupGEM mean-field) to further reduce the system-dimension, which now includes a moment-closure approximation.

In this chapter, we develop a continuous-time Markov model for the general group-based network model framework. This generalization covers any compartmental epidemic models, any static networks (e.g., directed, undirected, weighted), and any disjoint network partitions. Generalizations of a model can increase its flexibility, compatibility, and applicability. Different dynamics can be modeled with different compartmental models. In epidemiology, some disease dynamics can be modeled by the susceptible-infected-susceptible (SIS) or susceptible-infected-recovered (SIR) compartmental model, while some needed more complex compartmental models. A general framework can offer more flexibility to researchers to model epidemic processes on a network. The general group-based GroupGEM framework has lower computational complexity and faster simulation time in comparison with the general individual-based GEMF framework because of the reduced-state space size.

This chapter is organized as follows: the related knowledge of the work is reviewed in section 4.2 and 4.3. In section 4.4, we propose a continuous-time Markov process for a general group-based epidemic modeling framework (GroupGEM). Then we provide the mean-field approximation for this framework (GroupGEM mean-field) in section 4.5. We also provide some simulation results to compare between individual-based and group-based framework in synthetic networks and empirical networks. In section 4.6, we provide the multilayer extension of the GroupGEM framework. Last, we provide future directions (section 4.7) and concluding remarks on our work (section 4.8).

4.2 Compartmental epidemic models

Compartmental models can describe epidemic processes on a network $\mathcal{G}(N, E)$ [7]. Here, N is the number of nodes, and E is the set of edges in the network $\mathcal{G}$. In this chapter, we present a group-based framework for any compartmental epidemic model. We consider two types of transitions between compartments: 1) nodal transitions, and 2) edge transitions [1]. A nodal transition of a node depends on the current state of the node. An edge transition of a node depends on the current state of the node and the state of neighboring nodes. Each edge transition has an influencer compartment that is the compartment of the neighboring nodes or state of the neighboring nodes, which affect the edge transition. For example, in a susceptible-infected-susceptible epidemic, the susceptible-to-infected edge transition of a susceptible node is caused by its infected neighboring nodes. Therefore, the infected compartment is the influencer compartment for this edge transition. Even though our focus is on the group-based approach, these node-level transitions are still relevant. The change of a group-state occurs because of these node-level transitions. We call these transitions are events when an event (nodal or edge transition) happens on a node of a group, the group-state changes.

Prevalent epidemic compartmental models are:

4.2.1 Susceptible-infected-susceptible (SIS)

This model has two compartments, $m \in \{1, 2\}$: susceptible ($m = 1$) and infected ($m = 2$). A node in the network can be susceptible or infected. There are two types of transitions in this model: one is edge transition (susceptible-to-infected), and another is a nodal transition (infected-to-susceptible). A susceptible-to-infected transition of a node depends on the infected neighbors of that node. The infected compartment is the influencer compartment for the transition susceptible-to-infected. In GroupGEM, each group will have two types of nodes: susceptible and infected. The group-state will tell how many nodes are in each compartment. Let a group be in a state with S susceptible nodes and I infected nodes. If one infected node changes its compartment to the susceptible compartment, the group-state will change to a new state where it has $S + 1$ susceptible nodes and $I - 1$ infected nodes.

4.2.2 Susceptible-infected-recovered (SIR)

This model has three compartments, $m \in \{1, 2, 3\}$: susceptible ($m = 1$), infected ($m = 2$), and recovered ($m = 3$). Recovered nodes acquire immunity and cannot be infected anymore. Each group can have these three types of nodes. There are also two types of transitions in this model: susceptible-to-infected (edge transition) and infected-to-recovered (nodal transition). In GroupGEM, each group will have three types of nodes: susceptible, infected, and recovered.

4.2.3 Susceptible-exposed-infected-recovered (SEIR)

The SEIR model is a variation of the SIR model. This model has four compartments, $m \in \{1, 2, 3, 4\}$: susceptible ($m = 1$), exposed ($m = 2$), infected ($m = 3$), and recovered ($m = 4$). An exposed node is infected but not yet infectious. There are three types of transitions in this epidemic model: susceptible-to-exposed (edge transition), exposed-to-infected (nodal transition), and infected-to-recovered (nodal transition). For this model, the infected compartment is the influencer compartment for the susceptible-to-exposed edge transition.

These are basic, widely used epidemic compartmental models. A compartmental model can have any number of compartments and any number of transitions. Compartment number and type can be different in the scenario of rumor spreading or computer virus spreading.

4.3 Generalized epidemic modeling framework (GEMF) [1]

The GEMF is an individual-based continuous-time Markov epidemic modeling framework. A continuous-time Markov chain can model an epidemic process on a network when each transition between compartments is an independent Poisson process with a constant transition rate [117]. The assumption of the independent Poisson process makes the system memoryless.

In the individual-based GEMF process, nodes are at the individual level. Each node has a fixed number of possible states. The state of a node in a network for an M compartmental epidemic model at time t is defined as $n_i(t) \in 1, 2, 3, \ldots, M$. In an SIS epidemic process, if a node move from compartment 2 to compartment 1 in a Δt time by a nodal transition with rate δ, then waiting

time for this transition in GEMF is exponentially distributed with rate δ. So,

$$Pr[n_i(t + \Delta t) = 1 | n_i(t) = 2] = \delta \Delta t + o(\Delta t) \tag{4.1}$$

Here, $o(\Delta t)$ is a function of higher-order terms of Δt.

For an edge transition of node i from compartment 1 to 2 with a transition rate β, when the node has one infected neighbor, the infection process for the node i is

$$Pr[n_i(t + \Delta t) = 2 | n_i(t) = 1] = \beta \Delta t + o(\Delta t) \tag{4.2}$$

An infective link can transmit the disease with a constant rate β. In a network of N nodes, the individual-based continuous time Markov model GEMF has M^N possible states for an M compartmental epidemic model [1].

In the following section, we present a group-based continuous-time Markov process epidemic modeling on a network. All symbols and their definitions to develop this model are given in Table 4.1.

4.4 A group-based general epidemic modeling framework: GroupGEM

The first step to develop the group-based framework is to form the group-based adjacency matrix from the individual-based adjacency matrix, which describes the connections at the group-levels. A network consists of N nodes, which are divided into C disjoint non-empty groups. So

$$N = \mathcal{N}_1 + \mathcal{N}_2 + ... + \mathcal{N}_C \tag{4.3}$$

Here, $\mathcal{N}_i$ represents the number of nodes in a group i and $i = 1, 2, ..., C$. The adjacency matrix A of the network $\mathcal{G}$ is a $N \times N$ matrix, where each element is a binary number,

Table 4.1: Notation of parameters

Symbol	Definition
N	number of nodes
m	index variable for compartments
M	number of compartments in the epidemic model
t	time
n_i	state of the node i
C	no. of groups
L_{ij}	no. of links from group i to group j
$\mathcal{N}_i$	no. of nodes in group i
A	adjacency matrix (dimensions are $N \times N$)
$\mathcal{A}_g$	group-based adjacency matrix (dimensions are $C \times C$)
V_i	group-state matrix of a group i
$m \rightarrow n$	transitions of a node from compartment m to n
$k \rightarrow l$	transitions of a group from group-state k to l
$x_{i,m}$	no. of nodes in the compartment m in a group i
$x_{j,r}$	no. of nodes in the influencer compartment r in a group j. r is the influencer compartment for the edge transition $m \rightarrow n$
g_i	group-state of a group i
G	network-state
$\otimes$	kronecker product
θ_i	group transition rate matrix for a group i
Δ_{i,δ_q}	*transition-specific matrix* for group i for a nodal transition δ_q
Δ_{i,β_q}	*transition-specific matrix* for group i for an edge transition β_q
q_n	no. of nodal transitions
q_e	no. of edge transitions
Q	network-state transition matrix
ρ_i	fraction of the nodes in each compartment in group i
$\mathcal{L}$	no. of layers in the multilayer network
A_{gl}	group-based adjacency matrix for a layer l in the multilayer network

$$A(i,j) = \begin{cases} 1; & \text{if node } i \text{ is connected with } j \text{ by a link} \\ 0; & \text{otherwise} \end{cases} \tag{4.4}$$

The group-based adjacency matrix $\mathcal{A}_g$ of the network $\mathcal{G}$ for a C disjoint partitions is a $C \times C$ matrix. An element of the matrix $\mathcal{A}_g(i,j)$ represents the links from group i to group j,

$$\mathcal{A}_g(i,j) = \frac{\text{no. of links from group } i \text{ to group } j}{\mathcal{N}_i \mathcal{N}_j} = \frac{L_{ij}}{\mathcal{N}_i \mathcal{N}_j} \tag{4.5}$$

Here, L_{ij} indicates the number of links from group i to group j.

$$L_{ij} = u_i A u_j^T \tag{4.6}$$

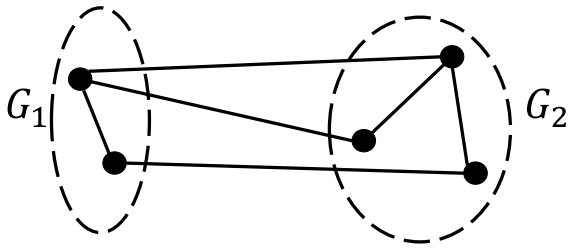
Where, u_i is a $1 \times N$ vector, if k^{th} node is in group i, then $u_i(k) = 1$ otherwise $u_i(k) = 0$.

The group-based adjacency matrix $\mathcal{A}_g$ is a symmetric matrix for an undirected network. Diagonal elements of the $\mathcal{A}_g$ matrix are $\frac{L_{ii}}{(\mathcal{N}_i)^2}$, where $L_{ii} = u_i A u_i^T$, so $diag(\mathcal{A}_g) \geq 0$. For the bipartite networks, $diag(\mathcal{A}_g) = 0$. If $C == N$, then $\mathcal{A}_g = A$. An example of a network with two groups is presented in Fig 4.1. The $\mathcal{A}_g$ matrix for this example is $\begin{bmatrix} \frac{L_{11}}{\mathcal{N}_1\mathcal{N}_1} & \frac{L_{12}}{\mathcal{N}_1\mathcal{N}_2} \\ \frac{L_{21}}{\mathcal{N}_2\mathcal{N}_1} & \frac{L_{22}}{\mathcal{N}_2\mathcal{N}_2} \end{bmatrix}$. The value of the $\mathcal{A}_g$ for the undirected network in Fig 4.1(a) is $\begin{bmatrix} \frac{2}{4} & \frac{3}{6} \\ \frac{3}{6} & \frac{4}{9} \end{bmatrix}$ and for the directed network in Fig 4.1(b) is $\begin{bmatrix} \frac{1}{4} & \frac{2}{6} \\ \frac{1}{6} & \frac{2}{9} \end{bmatrix}$.

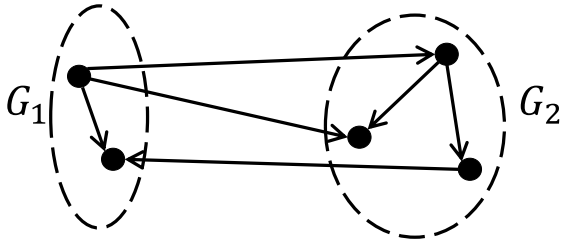
4.4.1 Group-state

The group-based model does not contain information about each node-state but includes the state of each group. The group-state $g_i(t)$ of a group i is a summary of its nodes state at time t. It reports how many nodes in the group i are in which compartments.

From the stars and bars combinatorics problem [118], the number of possible group-states of a group with $\mathcal{N}_i$ nodes for an M compartmental epidemic model is $\binom{\mathcal{N}_i+M-1}{M-1}$. The *stars and bars* combinatorics problem tells us the number of possible ways to put $\mathcal{N}_i$ indistinguishable nodes into M distinguishable compartments.



(a)



(b)

Figure 4.1: **A network with $N = 5$ nodes, which are divided into two groups, $C = 2$;** (a) undirected network, and (b) directed network.

Definition 1. ***(group-state matrix)*** *The group-state matrix V_i contains all the possible group-state of the group i. Each row of the V_i matrix represents each possible combination. The V_i of a group i is a $\binom{\mathcal{N}_i+M-1}{M-1} \times M$ matrix. Each row of the group-state matrix V_i is*

$$[x_{i,1}, x_{i,2}, x_{i,3}.....x_{i,M}] \tag{4.7}$$

Here, $x_{i,m}$ represents the number of nodes in the compartment m in the group i, also $0 \leq x_{i,1}, x_{i,2}, x_{i,3}....x_{i,M} \leq \mathcal{N}_i$ and $\sum_{m=1}^{m=M} x_{i,m} = \mathcal{N}_i$.

For example, $\mathcal{N}_i = 2, M = 3$, and number of possible group-state is $\binom{\mathcal{N}_i+M-1}{M-1} = 6$. Possible states for this case are

$$\overbrace{\begin{bmatrix} | & | & * & * \\ | & * & | & * \\ | & * & * & | \\ * & | & | & * \\ * & | & * & | \\ * & * & | & | \end{bmatrix}}^{\text{Dividers and nodes}} \qquad \overbrace{\begin{bmatrix} 0 & 0 & 2 \\ 0 & 1 & 1 \\ 0 & 2 & 0 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \\ 2 & 0 & 0 \end{bmatrix}}^{V_i} \tag{4.8}$$

Here, | represents a divider, and $*$ represents a node. Two dividers can divide the nodes into three compartments. The left matrix represents a chart of dividers and nodes. In each row, nodes on the left side of the first divider are in the first compartment, nodes between the first divider and the second divider are in the second compartment, and nodes on the right side of the second divider are in the third compartment. Each row represents a possible group-state. The first row represents that the first compartment has zero nodes, the second compartment has zero nodes, and the third compartment has two nodes. So, $x_{i,1} = 0, x_{i,2} = 0$, and $x_{i,3} = 2$. The right matrix presents the group-state matrix V_i, where the first column is the number of nodes in the first compartment $x_{i,1}$, the second column is the number of nodes in the second compartment $x_{i,2}$, and the third column is the number of nodes in the third compartment $x_{i,3}$.

Definition 2. ***(group-state indicator vector)*** *The group-state indicator vector e_i^k indicates that the group i is in the k^{th} possible state, e_i^k is a $\binom{N_i+M-1}{M-1} \times 1$ vector. If the group i is in the k^{th} possible state (details of the k^{th} possible state is in the k^{th} row of the group-state matrix V_i), then the k^{th} element of e_i^k is* 1*, and all the other elements are* 0*.*

We propose the pattern in expression (4.8) to organize the dividers and nodes. This pattern allows us to find the summary of nodes state of a group from the state indicator vector. If $g_i(t) = e_i^k$, then the group i is in the k^{th} possible state at time t. The group-state $g_i(t)$ is a vector with $\binom{N_i+M-1}{M-1}$ elements where all elements are zero except one element corresponding to the group-state.

4.4.2 Network-state

Network-state $G(t)$ is the joint-state of all groups at a time t. The network-state $G(t)$ can be defined as

$$G(t) = g_1(t) \otimes g_2(t) \otimes ..... \otimes g_C(t) \tag{4.9}$$

Here, $\otimes$ represents the Kronecker product, which allows one-hot encoding to store the state space and the infinitesimal generator matrix of the Markov chain [119, 120]. The dimensions of the joint-state vector $G(t)$ are $\left[\prod_{i=1}^{C}\binom{N_i+M-1}{M-1}\right] \times 1$. In $G(t)$, all elements are zero except one element corresponding to the network-state.

The disease propagation process of an M compartmental epidemic model in the network with N nodes forms a continuous time Markov process. The number of states in the Markov chain or the number of possible network-state in the group-based framework is $\prod_{i=1}^{C}\binom{N_i+M-1}{M-1}$.

The number of states in the group-based Markov model is less than or equal to the number of states in the individual-based Markov model. Therefore, $\prod_{i=1}^{C}\binom{N_i+M-1}{M-1} \leq M^N$. If $C = N$, then $\prod_{i=1}^{C}\binom{N_i+M-1}{M-1} = M^N$ which leads the group-based approach to the individual-based approach.

From the network-state $G(t)$, it is possible to infer each group-state $g_i(t)$ in the following way

$$g_i(t) = \left(1^T_{\binom{N_1+M-1}{M-1}\times 1} \otimes \cdots \otimes I_{\binom{N_i+M-1}{M-1}\times\binom{N_i+M-1}{M-1}} \otimes \cdots \otimes 1^T_{\binom{N_C+M-1}{M-1}\times 1}\right) G(t) \tag{4.10}$$

4.4.3 Group-level transitions

All events or transitions are modeled here as independent Poisson processes; therefore, waiting times for events are exponentially distributed. Hence, the system has the memoryless property. An event or transition in a group changes the network-state, which is the state transition in the Markov chain. Transitions in the group-state are also two types: 1) nodal transition, and 2) edge transition.

Nodal transition

This transition depends only on the state of a group. It does not depend on the state of its neighboring groups. If a nodal transition from compartment m to compartment n happens in a node of

group i with a rate δ_q that means a node in group i moves from compartment m to n. Therefore, the group will change its state for this nodal transition from group-state k to l with the rate $x_{i,m}\delta_q$. If group-state k has $x_{i,m}$ nodes in compartment m and $x_{i,n}$ nodes in compartment n, then group-state l has $x_{i,m} - 1$ nodes in compartment m and $x_{i,n} + 1$ nodes in compartment n. An example of the group-level transition of a group i for an infected-to-susceptible nodal transition in the SIS epidemic model is presented in Fig 4.2b.

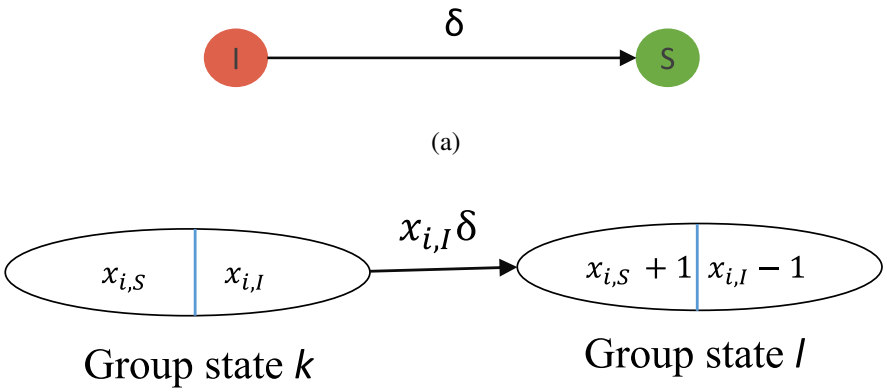


Figure 4.2: **An infected-to-susceptible nodal transition in the SIS epidemic model.** The arrow lines indicate the transition. (a) Node-level transition: an infected node in the group i moves from infected to susceptible compartment with δ rate, (b) group-level transition: group i moves from group state k to l with $x_{i,I}\delta$ rate for the infected-to-susceptible nodal transition of a node in the group i. Group state k has $x_{i,S}$ susceptible nodes and $x_{i,I}$ infected nodes. Group state l has $x_{i,S} + 1$ susceptible nodes and $x_{i,I} - 1$ infected nodes. This figure only presents transition for an infected-to-susceptible nodal transition; the other possible transitions are not visible here.

Edge transition and an approximation

An edge transition of a "group" i depends on its own state and the state of its neighboring groups. An edge transition of a "node" in a group i depends on its neighboring nodes, which are distributed in different groups. As an example, the susceptible-to-infected edge transition in a group i happens if a susceptible node in group i is in contact with at least one infected node. However, the group-level framework does not contain information about which node is in which compartment. Also, the group-level adjacency matrix cannot tell about the exact neighbors of a node. Therefore, the edge transition at the group-level needs an approximation at the network-level.

The rate of an edge transition ($m \to n$) depends on the number of edges from the nodes of compartment m to the nodes of influencer compartment r for that edge transition. Let $\mathcal{X}$ be the set of nodes that are in compartment m in group i, and $\mathcal{Y}$ is the set of nodes that are in compartment r in the group j. Now, $L_{\mathcal{X}\mathcal{Y}}$ is the number of edges between $\mathcal{X}$ and $\mathcal{Y}$. The red dotted lines in Fig 4.3 represent $L_{\mathcal{X}\mathcal{Y}}$, where m is the susceptible compartment, n is the infected compartment, and r is the infected compartment in the SIS dynamic (here, $n = r$). The infection rate in group i for the group j is proportional to $L_{\mathcal{X}\mathcal{Y}}$ (red dotted links in the Fig 4.3). The group-level aggregation keeps the track

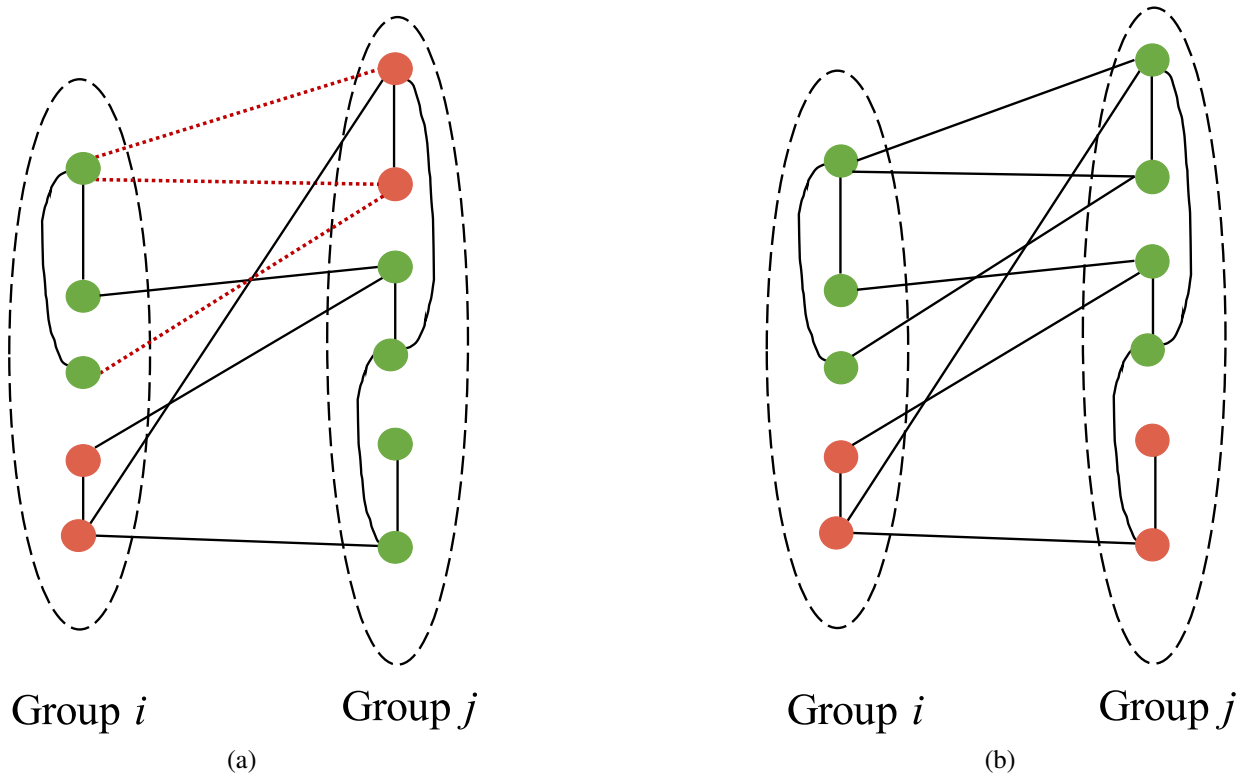


Figure 4.3: **A graphical explanation of the topological approximation.** A network with two groups is presented for a two compartmental epidemic model (susceptible-infected-susceptible SIS). The green nodes are in the compartment m, and the red nodes are in the influencer compartment r. The m compartment is the susceptible compartment, and the influencer r compartment is the infected compartment. The red dotted edges represent the edges from susceptible nodes in group i to infected nodes in group j ($L_{\mathcal{X}\mathcal{Y}}$ in Eq. 4.12 is the number of such red dotted edges). Sub-figure (a), and (b) are two possible combination of infected nodes in group j. Here, $L_{i,j} = 7$, $\mathcal{N}_i = 5$, $\mathcal{N}_j = 6$, $x_{i,m} = 3$, $x_{j,r} = 2$. Therefore, $\mathcal{A}_g(i,j)x_{i,m}x_{j,r} = 1.4$. In the sub-figure (a), $L_{\mathcal{X}\mathcal{Y}} = 3$, and in the sub-figure (b), $L_{\mathcal{X}\mathcal{Y}} = 0$. The infection rate in group i for group j is proportional to $\mathcal{A}_g(i,j)x_{i,m}x_{j,r}$ in the group-based approach. Therefore, $\mathcal{A}_g(i,j)x_{i,m}x_{j,r}$ approximates the expected of the different possible combinations.

of number of node in each compartment in a group and number of edges between two groups. For the edge-based transition, the group-level aggregation needs an topological approximation, which is defined as [28]

$$L_{\mathcal{XY}} \approx \mathcal{A}_g(i,j)x_{i,m}x_{j,r} \tag{4.11}$$

Here, $\mathcal{A}_g(i,j) = \frac{L_{ij}}{N_i N_j}$; L_{ij} is the number of edges between group i and j; $x_{i,m}$ is the number of nodes in the compartment m in group i; and $x_{j,r}$ is the number of nodes in the influencer compartment r in group j.

It is possible to give a bound on the topological approximation from the discrete isoperimetric inequality [2] (Eq. 24 in [28]). Suppose that a network has N nodes which are divided into C disjoint subsets. The discrete isoperimetric inequality [28, 121] for the number of links $L_{\mathcal{XY}}$ with the first endpoint in subset $\mathcal{X}$ and the second endpoint in subset $\mathcal{Y}$ is

$$|L_{\mathcal{XY}} - \mathcal{A}_g(i,j)x_{i,m}x_{j,r}| \leq \frac{\theta}{N}\sqrt{x_{i,m}(N-x_{i,m})x_{j,r}(N-x_{j,r})} \tag{4.12}$$

Here, θ is $\theta \in R$ and $|d - \sigma_i| \leq \theta$ for $i \neq 0$. d is the average node degree, σ_i are the eigenvalues of the Laplacian matrix of the network for $1 \leq i < N$.

There are other ways to give tighter bounds on this approximation derived from the Max-Cut problem and the expander mixing lemma [122].

An edge transition of a node depends on its neighborhood. However, the group-based adjacency matrix does not contain information about the exact neighborhood of a node. Therefore, this group-based model requires a topological approximation (Eq. 4.11), which is bounded by the isoperimetric inequality (Eq. 4.12). The Szemerédi's regularity lemma [123] may give an intuition for which partitions in a network the topological approximation will be less erroneous. This regularity lemma is a powerful tool in the extremal graph theory. According to this lemma, the nodes in a large enough network can be grouped into a bounded number of groups so that the edges between different groups behave almost randomly.

[2] Isoperimetric inequality is an ancient Greek problem. The isoperimetric inequality is a geometric problem - finding the closed curve among all the possible curves of a given perimeter, which encloses the maximum area [28, 121].

4.4.4 Evolution of the network-state

The evolution of the network-state $G(t)$ follows a continuous-time Markov process. The network-state is the joint-state of all group-states. In the group-based framework, groups are interacting entities, which are jointly Markovian and form a collective system.

In the individual-based approach, each node is different concerning its connections with the neighboring set of nodes; however, each node has the same set of possible states. In the group-based approach, each group is also different concerning its connections with the neighboring set of groups; however, this approach has another complexity that each group can have different sets of possible group-states corresponding to their sizes. For example, in an SIS epidemic, each node in the individual-based approach has two possible states: susceptible and infected. However, in the group-based approach, each group has a different number of possible states dependent on their size (number of nodes in that group); the number of possible states in a bigger group is larger than in a smaller group. Therefore, the group-based Markovian process has another level of complexity beyond the individual-based one.

Transition Rate

The group-state of a group changes when a node in the group changes its compartment. Waiting time for any event or transition in the network is stochastically independent.

Definition 3. ***(Transition-specific Matrix)*** *A transition-specific matrix of a group i for a transition $m \rightarrow n$ indicates the group-state changes for the transition $m \rightarrow n$. Transition-specific matrices are group specific and transition specific. Transition-specific matrices are two types: 1) nodal transition-specific matrix Δ_{i,δ_q}, and 2) edge transition-specific matrix Δ_{i,β_q}. If a node moves from m compartment to n compartment for the q-type nodal transition and the group-state changes from k*

to l, then the elements of $\Delta_{\delta_q i}$ can be defined as

$$\Delta_{i,\delta_q}(\mathcal{I},\mathcal{J}) = \begin{cases} \delta_q V_i(k,m) & \text{if } \mathcal{I} = k \text{ and } \mathcal{J} = l \\ -\delta_q V_i(k,m) & \text{if } \mathcal{I} = k \text{ and } \mathcal{J} = k \\ 0 & \text{otherwise} \end{cases} \tag{4.13}$$

Here, $\mathcal{I}$ and $\mathcal{J}$ are index variables for the matrix Δ_{i,δ_q}. The definition of Δ_{i,β_q} is the same as Eq. 4.13.

If an epidemic model has q_n nodal transitions and q_e edge transitions, then the framework has $q_n + q_e$ transition-specific matrices for a group i. For example, in the SIR epidemic model $q_n = 1$ and $q_e = 1$; in the SEIR epidemic model $q_n = 2$, and $q_e = 1$.

Definition 4. ***(Group Transition Rate Matrix)*** *The group transition rate matrix θ_i of a group i stores all the possible transitions rate of the group i. The $\theta_i(k,l)$ element of the group transition rate matrix will store the transition rate from group-state k to l of a group i. If a group has $\binom{N_i+M-1}{M-1}$ possible states then the dimensions of the group transition rate matrix are $\binom{N_i+M-1}{M-1} \times \binom{N_i+M-1}{M-1}$.*

If an epidemic model has q_n types of nodal transitions and q_e types of edge transitions, then an element of the group transition rate matrix $\theta_i(k,l)$ has $q_n + q_e$ parts. This element represents the transition from state k to l, and each part corresponds to each transition. A part of $\theta_i(k,l)$ for a q type nodal transition will be

$$(e_i^l)^T \Delta_{i,\delta_q}^T e_i^k \tag{4.14}$$

Now, let us consider a q-type edge transition from compartment m to n for the influencer compartment r with the rate β_q, which depends on the r compartmental neighboring nodes. The group-state transition rate of a group i for the edge transition β_q is

$$\beta_q \sum_{j=1}^{j=C} x_{i,m} \mathcal{A}_g(i,j) x_{j,r} \tag{4.15}$$

Here, $x_{i,m}$ is the number of nodes in the m compartment in the group i and $x_{i,m} = (V_i(:,m))^T g_i(t)$. Now, we can define the part of $\theta_i(k,l)$ for this edge transition as,

$$(e_i^l)^T \sum_{j=1}^{C} \{\mathcal{A}_g(i,j) x_{j,r}\} \Delta_{i,\beta_q}^T e_i^k \tag{4.16}$$

Therefore, an element of the group transition rate matrix θ_i is

$$\theta_i(k,l) = \sum_{q=1}^{q_n} (e_i^l)^T \Delta_{i,\delta_q}^T e_i^k + \sum_{q=1}^{q_e} (e_i^l)^T \Big(\sum_{j=1}^{C} \mathcal{A}_g(i,j) x_{j,r}\Big) \Delta_{i,\beta_q}^T e_i^k \tag{4.17}$$

Here, $\theta_i(k,l)$ represents the rate for the group transition from state k to l. The Eq. (4.17) has two parts; the first is for all types of nodal transitions, and the second is for all kinds of edge transitions. The group transition rate matrix θ_i is group-specific; different groups will have different θ_i matrices.

Group-based Markov process

The evolution of the network-state $G(t)$ is a continuous-time Markov process, where actual Markov states are the possible network-states. This process is fully characterized by systems of differential equations named as the Kolmogorov differential equations for a given initial condition. The procedure to derive the Kolmogorov differential equations for a Markov chain from one-state transition rates is described in ref. [124, 125].

Theorem 1. *The Kolmogorov differential equations describe the dynamics of the underlying group-based continuous-time Markov process for an epidemic process in a network with N nodes, which are divided into C disjoint subsets. The Kolmogorov differential equations for the group-based framework are*

$$\frac{d}{dt} E[G] = Q^T E[G] \tag{4.18}$$

Eq. (4.18) is the time evolution of the network-state. Here, Q is the infinitesimal generator of the underlying Markov process (computed in the Appendix B.1). This closed set of differential

equations can fully characterize the network-state. Dimensions of the Q are $[\prod_{i=1}^{C}\binom{N_i+M-1}{M-1}] \times [\prod_{i=1}^{C}\binom{N_i+M-1}{M-1}]$.

Proof. If a network is in the $G(t)$ state at time t, then the evolution of the network-state tells us the network-state after Δt time $G(t+\Delta t)$. To obtain $G(t+\Delta t)$, we first derive the expression for the state of group i at time $(t+\Delta t)$, given that the network is in the $G(t)$ state at time t, which is $Pr[g_i(t+\Delta t) = e_i^l | g_i(t) = e_i^k, G(t)]$. This expression indicates the probability of a transition of group i from state k to l in time interval $(t, t+\Delta t]$.

The transition from group-state k to l is an independent Poisson process, which occurs in $(t, t+\Delta t]$ time interval. Therefore,

$$Pr[g_i(t+\Delta t) = e_i^l | g_i(t) = e_i^k, G(t)] = \theta_i(k,l)\Delta t + o(\Delta t) \tag{4.19}$$

Eq. (4.19) will be used to derive the time evolution of the network-state. The group transition rate matrix θ_i gives us a description of the group-level transition.

The expected value of a group-state in the next time step, when the network is in the $G(t)$ state, can be obtained from Eq. (4.17) and (4.19),

$$E[g_i(t+\Delta t)|G(t)] = \sum_{q=1}^{q_n} \Delta_{i,\delta_q}^T g_i(t)\Delta t + \sum_{q=1}^{q_e}(\sum_{j=1}^{C} \mathcal{A}_g(i,j)x_{j,r})\Delta_{i,\beta_q}^T g_i(t)\Delta t + g_i(t) + o(\Delta t) \tag{4.20}$$

Now, considering the expected value of both sides in Eq. (4.20),

$$E[E[g_i(t+\Delta t)|G(t)]] = E[g_i(t+\Delta t)] = \sum_{q=1}^{q_n} \Delta_{i,\delta_q}^T E[g_i(t)]\Delta t + \sum_{q=1}^{q_e} \Delta_{i,\beta_q}^T E[h_i(t)g_i(t)]\Delta t + E[g_i(t)] + E[o(\Delta t)] \tag{4.21}$$

In Eq. (4.21), the expression for $h_i(t)$ is

$$h_i(t) = \sum_{j=1}^{C} \mathcal{A}_g(i,j)x_{j,r} \tag{4.22}$$

Here, compartment r is the influencer compartment for the q^{th} edge transition.

After rearranging the Eq. (4.21) as follows:

$$\frac{E[g_i(t+\Delta t)] - E[g_i(t)]}{\Delta t} = \sum_{q=1}^{q_n} \Delta_{i,\delta_q}^T E[g_i(t)] + \sum_{q=1}^{q_e} \Delta_{i,\beta_q}^T E[h_i(t) g_i(t)] + \frac{E[o(\Delta t)]}{\Delta t} \quad (4.23)$$

Let $\Delta t \to 0$, so the Eq. (4.23) will become,

$$\frac{d}{dt} E[g_i(t)] = \sum_{q=1}^{q_n} \Delta_{i,\delta_q}^T E[g_i(t)] + \sum_{q=1}^{q_e} \Delta_{i,\beta_q}^T E[h_i(t) g_i(t)] \quad (4.24)$$

The rest of the proof is given in Appendix B.1. □

The network-state evolution in the group-based structure is a multidimensional birth-death process, a particular case of the continuous-time Markov process. Numerical simulation of the collective network system is useful in understanding the system dynamics.

Remark. *By partitioning the population into C groups, we go from M^N states to $\binom{\frac{N}{C}+M-1}{M-1}^C$ states, which is already polynomial in N for a constant number of groups and quasi-polynomial in N for a logarithmic number of group, i.e., $C = O(\log N)$.*

4.5 Mean-field approximations of the GroupGEM

It is possible to reduce the state-space size of the continuous-time Markov model of the group-based approach (Eq. 4.18) by using closure approximations techniques. In this chapter, we propose two levels of first-order moment closure approximation to further reduce the state space size: 1) inter-group mean-field approximation, 2) intra- and inter-group mean-field approximation.

4.5.1 Inter-group mean-field approximation

In the inter-group mean-field approximation, we assume states of a group are uncorrelated with other groups, which gives us the joint probability distribution of states within groups.

Theorem 2. *The inter-group first-order mean-field approximation reduces the state space size of the group-based Markov model to* $\sum_{i=1}^{C}\binom{N_i+M-1}{M-1}$. *The inter-group mean-field equation of the group-based framework is*

$$\frac{d}{dt}E[g_i(t)] \quad = \quad \sum_{q=1}^{q_n}\Delta_{i,\delta_q}^{T}E[g_i(t)] \quad + \quad \sum_{q=1}^{q_e}\Big(\sum_{j=1}^{C}\mathcal{A}_g(i,j)(V_j^g(:,r))^T E[g_j(t)]\Big)\Delta_{i,\beta_q}^{T}E[g_i(t)] \quad (4.25)$$

Proof. The Eq. (4.24) and (4.18) contain higher-order moment terms $E[h_i(t)g_i(t)]$. This framework has assumed that states of individual groups are independent random variables and invoke moment-closure approximation for those higher-order moments. This approximation allows us to assume the covariance between two random variable $h_i(t)$ and $g_i(t)$ is zero. From the first moment-closure approximation we can write,

$$Cov[h_i(t)g_i(t)] \approx 0 \quad (4.26)$$

$$\Rightarrow \quad E[h_i(t)g_i(t)] \quad \approx \quad E[h_i(t)]E[g_i(t)] \quad \approx \quad \sum_{j=1}^{C}\mathcal{A}_g(i,j)(V_j(:,r))^T E[g_j(t)]E[g_i(t)] \quad (4.27)$$

□

4.5.2 Intra- and inter-group mean-field approximation

The intra- and inter-group mean-field approximation assumes uncorrelation within and across the groups. If a node in a group i moves its compartment from m to n with a rate of δ_q, then the population in the corresponding compartments will change in the group i (Fig 4.2). The nodal transition matrix Φ_{δ_q} for the intra- and inter-group mean-field represents a nodal transition from compartment m to n with rate δ_q. It has dimensions $M\times M$, where $\Phi_{\delta_q}(m,m) = -\delta_q$, $\Phi_{\delta_q}(m,n) = \delta_q$, and otherwise zero. This matrix has the form of a Laplacian matrix. An edge transition matrix Φ_{β_q} has similar structure.

Theorem 3. *The intra- and inter-group mean-field equation for the group-based framework is,*

$$\frac{d}{dt}E[X_i] = \sum_{q=1}^{q_n} \Phi_{\delta_q}^T E[X_i] + \sum_{q=1}^{q_e} \Big(\sum_{j=1}^{C} \mathcal{A}_g(i,j) E[X_{j,r}] \Big) \Phi_{\beta_q}^T E[X_i] \quad (4.28)$$

Here, $X_i = [x_{i,1}, x_{i,2}, x_{i,3}....., x_{i,M}]^T$.

The intra- and inter-group first-order mean-field approximation provides marginal state probabilities of each group yielding only MC equations, which is independent of the network size N for a constant number of groups C.

Proof. Appendix B.2. □

If the fraction of nodes in each compartment in group i is $\rho_i = \frac{E[X_i]}{\mathcal{N}_i}$, where $\rho_i = [\rho_{i,1}, \rho_{i,2}....., \rho_{i,M}]^T$ and $\sum_{m=1}^{M} \rho_{i,m} = 1$, then Eq. (4.28) can be written as

$$\dot{\rho_i} = \sum_{q=1}^{q_n} \Phi_{\delta_q}^T \rho_i + \sum_{q=1}^{q_e} \Big(\sum_{j=1}^{C} \frac{L_{ij}}{\mathcal{N}_i} \rho_{j,r} \Big) \Phi_{\beta_q}^T \rho_i \quad (4.29)$$

Group-based intra- and inter-group mean-field equations for SIS, SIR, and SEIR epidemic models are given in Appendix B.3.

4.5.3 Numerical experiments

We perform a numerical study to evaluate the performance of the group-based approach. We compare the simulation results of the intra- and inter-group mean-field equations with the exact Markov process of the individual-based approach, which does not have any approximation errors. We investigate the global and local dynamics of the Barabási-Albert ($N = 10000, m = 40$) scale-free random network, Erdös-Rényi ($N = 10000, p = 0.01$) random network, and two empirical networks. We also propose four different ways to group the nodes and investigate their performance.

Simulations on synthetic networks

The simulation results on a Barabási-Albert ($N = 10000, m = 40$) random network for an SIS ($\beta = 0.0167$ and $\delta = 1$), and an SIR ($\beta = 0.0167$ and $\delta = 1$) epidemic model are presented in Fig 4.4, 4.5, 4.6 and 4.7. The number of edges and average node degrees of this network are 399157 and 79.83, respectively. As an initial condition, we have started the epidemic by setting the first 0.2% of nodes as infected nodes.

The sub-figure (a) in Fig 4.4, and 4.6 represents the normalized population in different compart-

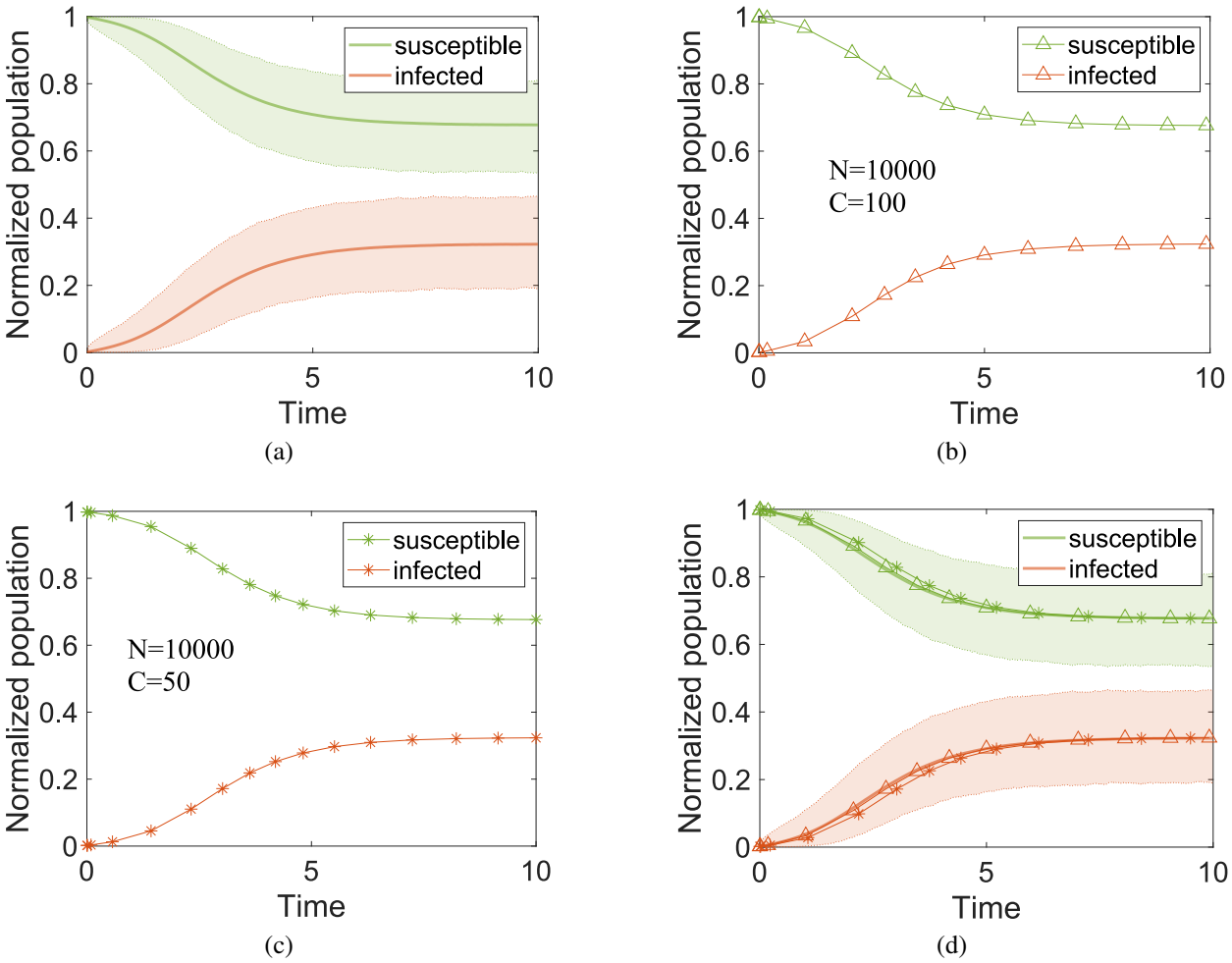


Figure 4.4: **Global dynamics of an SIS epidemic in a Barabási-Albert network** ($N = 10000, m = 40$)**;** a) stochastic numerical simulation of the individual-based continuous-time exact Markov model GEMF, solid lines represent the average of the 1000 simulations and shaded areas represent the region of the stochastic simulations; b) group-based: $C = 100, \mathcal{N}_1 = \mathcal{N}_2 = ... = \mathcal{N}_C = 100$, simulation time = $0.151s$; c) group-based: $C = 50, \mathcal{N}_1 = \mathcal{N}_2 = ... = \mathcal{N}_C = 200$, simulation time = $0.049s$; and d) merging of all sub-plots a-c.

ments of stochastic numerical simulations of the individual-based continuous-time exact Markov model GEMF. The simulation result of intra- and inter-group mean-field group-based framework is presented in sub-figure (b), and (c). In the sub-figure (b), nodes are divided into 100 groups sequentially, while each group has 100 nodes. Similarly, nodes are divided into 50 groups in the sub-figure (c), while each group has 200 nodes. Sub-figure (d) is presenting the merging of sub-plots (a)-(c).

The GroupGEM framework allows us to do the local dynamic analysis at the group-level. We present the local dynamic analyses in Fig 4.5 and 4.7. To see the local dynamics, we choose a partition, where nodes are divided into 100 groups sequentially (Fig 4.4(b), and 4.6(b)). At $t = 0$, first 20 nodes are infected, which are belongs to group 1. Fig 4.5(a) is presenting normalized infected populations in groups $1, 7$, and 63 in the SIS epidemic model. We have picked group 7 and 63 randomly. Fig 4.7(a) presents normalized infected populations in groups $1, 15$, and 92 in the SIR epidemic model. We average the state of all nodes of a group of the individual based stochastic simulations to compare the dynamic of a group with the group-based approach. Different groups have different levels of infected populations, which are also other than the global infected population. The GroupGEM is informative at the group-level for an epidemic. Fig 4.5(b) and 4.7(b) present the histogram of the absolute error for all groups at a fixed time.

A comparison of simulation time between the individual-based and group-based approaches of Fig 4.4 and 4.6 is given in Table 4.2. The computational environment was the same for each case. From Fig 4.4(d) and 4.6(d), the group-based approaches can produce similar dynamics as the individual-based approaches in SIS and SIR disease models in a Barabási-Albert network. From Table 4.2, the simulation time for group-based methods is less than the simulation time for the individual-based approaches. A similar investigation on an Erdös-Rényi random network is provided in Appendix B.4.

Simulations on empirical networks

We use GroupGEM intra- and inter-group mean-field approach to simulate susceptible-infected-susceptible (SIS) epidemic model ($\beta = 0.25, \delta = 1$, first 20 nodes are infected at $t = 0$) on two

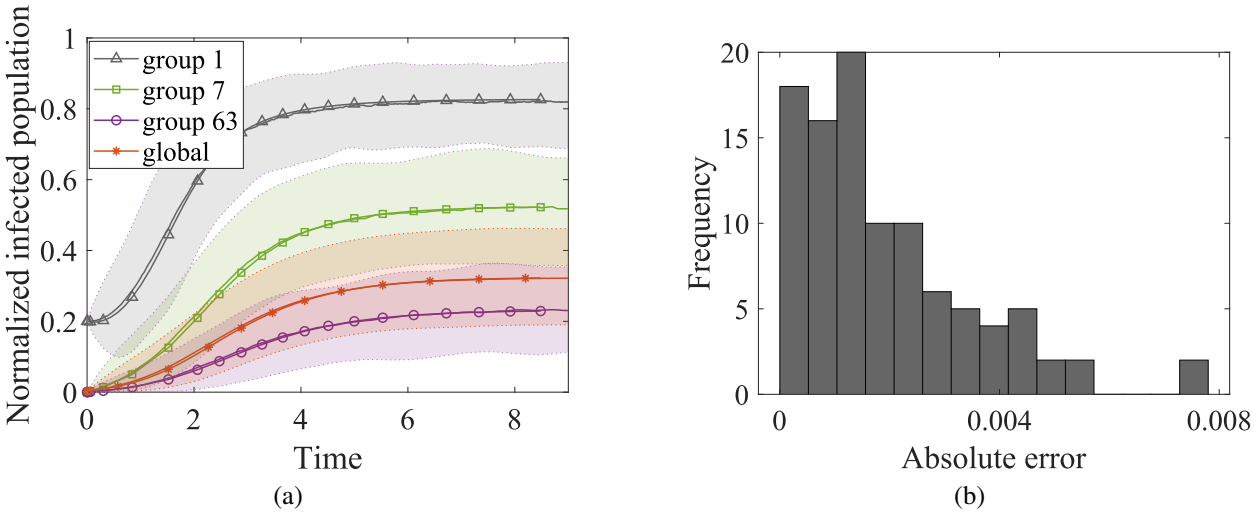


Figure 4.5: **Local dynamics at the group level for the case in sub-figure 4.4(b).** Simulations on a Barabási-Albert network ($N = 10000, m = 40$) (SIS epidemic model: $\beta = 0.25, \delta = 1$, node: $1-20$ were infected at $t = 0$). a) Time dynamics of the normalized infected population of group 1, 7, and 63. Solid lines represent the mean of stochastic individual-based simulation (average the dynamic of nodes of a group to compare), and shaded regions represent 1000 stochastic simulations. Solid lines with markers represent the output from the intra- and inter-group mean-field model. b) Histogram of absolute error of the normalized infected population of all groups at $t = 8$. Here, x-axis is the absolute error of the intra- and inter-group mean-field model compares to the mean of the stochastic individual-based simulations, and y-axis is the no. of the groups or frequency.

empirical social networks. The first network was generated from email communication between members of a large European research institution [2]. This network has 1005 nodes, and 25571 edges. The clustering coefficient of this network is 0.4. We divide this network into 56 groups, which reduces the simulation time more than 11 times compared to individual-based mean-field and more than 300 times compared to stochastic simulations (200 realizations). The output of the group-based approach can provide information about the group-states of this 56 group.

The second network was generated from links between users of Slashdot (a technology-related news website). It has 82168 nodes and 948464 edges [2]. We divide the nodes of this Slashdot network into 11914 groups. The group-based approach reduces simulation time more than 50 times compared to individual-based mean-field and more than 1100 compared to stochastic simulations (200 realizations). At any time t, the group-based model can provide the summary of nodes state of 11914 groups. We compare our results with an average of 200 stochastic simulations in Fig 4.8. The networks and the groupings are given in our publicly shared supporting dataset (http://ieee-

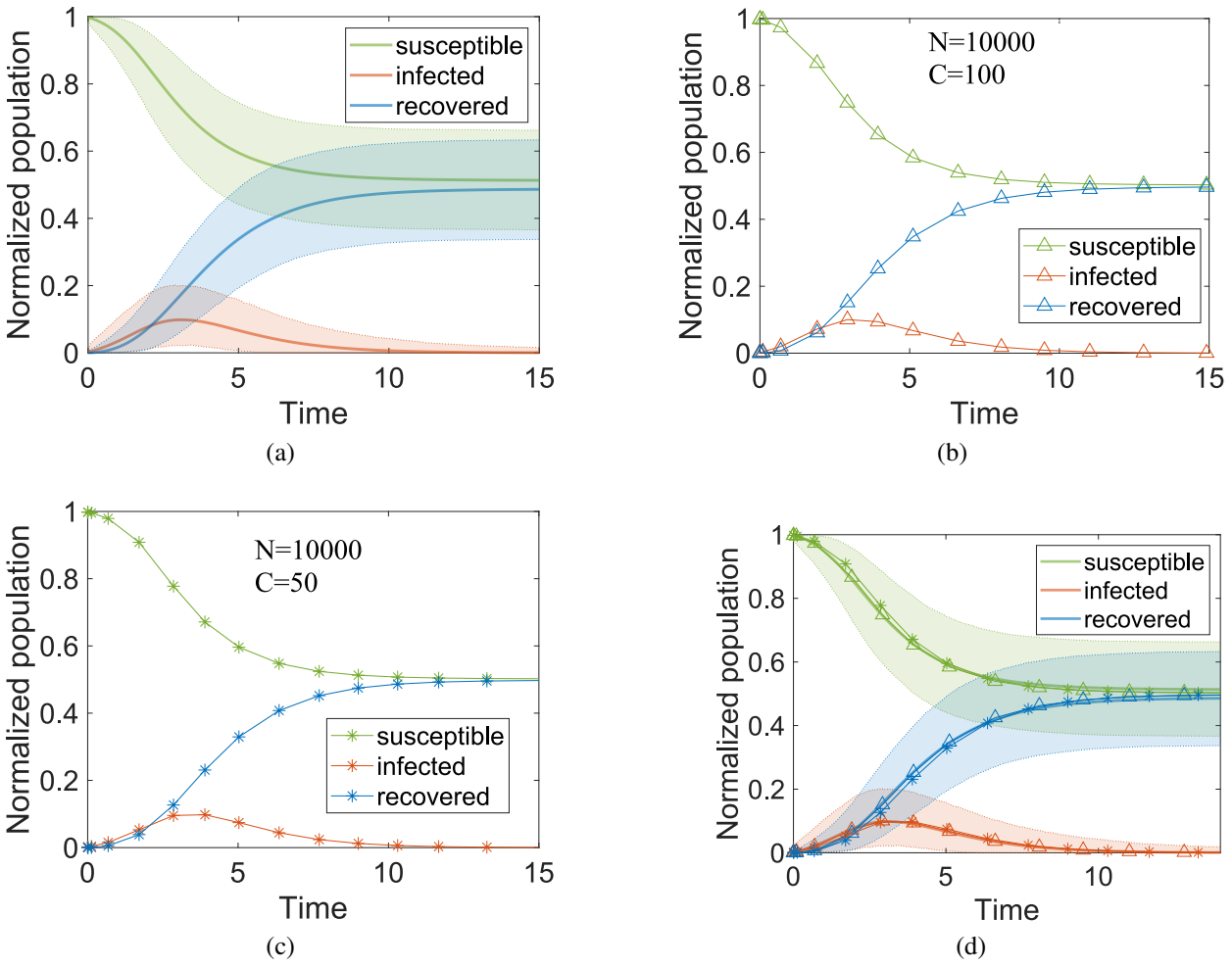


Figure 4.6: **Global dynamics of an SIR epidemic in a Barabási-Albert network** ($N = 10000, m = 40$)**;** a) stochastic numerical simulation of the individual-based continuous-time exact Markov model GEMF, solid lines represent the average of the 1000 simulations and shaded areas represent the region of the stochastic simulations; b) group-based: $C = 100, \mathcal{N}_1 = \mathcal{N}_2 = ... = \mathcal{N}_C = 100$, simulation time = $0.0.088s$; c) group-based: $C = 50, \mathcal{N}_1 = \mathcal{N}_2 = ... = \mathcal{N}_C = 200$, simulation time = $0.042s$; and d) merging of all sub-plots a-c.

dataport.org/2182).

The group-based GroupGEM framework allows us to do the local dynamic analysis at the group level. The GroupGEM framework gives the flexibility to start the epidemic from specific nodes. Local dynamic analysis for the email-Eu-core network [2] is presented in Fig. 4.9. At t= 0, node: 1 – 20 are infected and they are belongs to group 1. To see the time dynamics of the infected population at the group-level, we pick group 5 and 30 randomly (Fig. 4.9(a)). We use the average of the state of all nodes of a group of the individual based stochastic simulations to compare

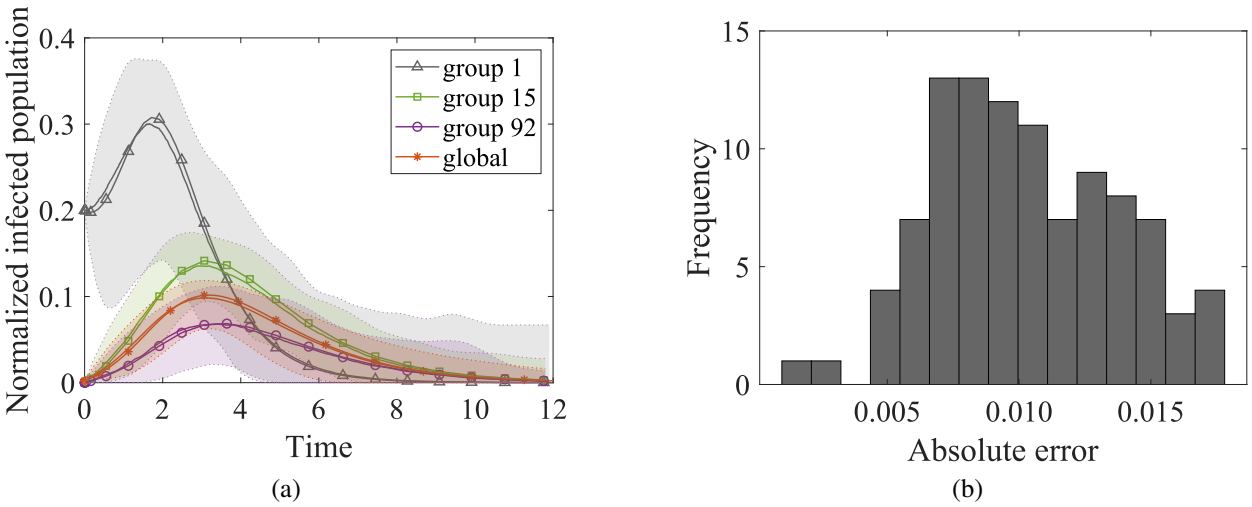


Figure 4.7: **Local dynamics at the group level for the case in sub-figure 4.6(b).** Simulations on a Barabási-Albert network ($N = 10000, m = 40$) (SIR epidemic model: $\beta = 0.25, \delta = 1$, node: $1 - 20$ were infected at $t = 0$). a) Time dynamics of the normalized infected population of group 1, 15, and 92. Solid lines represent the mean of stochastic individual-based simulation (average the dynamic of nodes of a group to compare), and shaded regions represent 1000 stochastic simulations. Solid lines with markers represent the output from the intra- and inter-group mean-field model. b) Histogram of absolute error of the normalized recovered population of all groups at $t = 12$. Here, x-axis is the absolute error of the intra- and inter-group mean-field model compares to the mean of the stochastic individual-based simulations, and y-axis is the no. of the groups or frequency.

the dynamic of a group with the group-based approach. The network and grouping of nodes are uploaded as a supplementary file in the IEEE data port (http://ieee-dataport.org/2182).

Grouping approaches

We investigate four conventional and intuitive heuristic ways to group the nodes: 1) random, 2) degree-based, 3) community, and 4) k-partite. In a random grouping, we randomly divide the nodes into different groups. In a degree-based grouping, nodes are grouped according to their degree. Nodes in a group have the same degree. The community grouping approach is only applicable to the networks which have community structure. In the community structure, nodes have more connections inside the communities than between communities. In a community grouping, we divide the nodes according to their community. The nodes of a group are from the same com-

Table 4.2: A comparison of simulation time between individual-based and group-based approaches.

case	No. of Groups	simulation time		
		SIS	SIR	SEIR
Individual-based stochastic GEMF (1000 realizations)	-	3140s	448s	895s
Individual-based mean-field	-	62.409s	12.183s	15.264s
group-based mean-field	100	0.151s	0.088s	0.147s
group-based mean-field	50	0.049s	0.042s	0.066s
group-based mean-field	10	0.016s	0.018s	0.019s

munity. In the k-partite grouping, we select groups in a way that a group does not have any internal connections. To choose a k-partite partition, we use a solution of the graph coloring problem [126].

We apply these four heuristic ways in four different networks: a) Barabási-Albert (BA) scale-free ($N = 10000, m = 40$) random network, b) Erdös-Rényi (ER) ($N = 10000, p = 0.01$) random network, c) stochastic-block model (SBM) ($N = 10000, pi = 0.05, po = 0.0001$) network, and d) email-Eu-core empirical network [2]. The first three synthetic networks have the same average node degree. Barabási-Albert (BA) and Erdös-Rényi (ER) network do not form any community structure; therefore, we do not inspect community grouping in these networks. We use a community grouping in the stochastic-block network and the email-Eu-core empirical network. This Stochastic-block network has 10 communities and sizes of the communities are 2000, 1000, 500, 600, 100, 2000, 300, 2000, 100, and 1400. The email-Eu-core network has 42 communities, which are defined from data [2].

Fig 4.10 presents the simulation results for the four heuristic grouping approaches. Fig 4.10 has a double y-axis: the left y-axis shows the simulation times, and the right y-axis presents the absolute error of the intra- and inter-group mean-field model. We calculate the absolute error of the metastable [14] infected population in the susceptible-infected-susceptible (SIS) epidemic process ($\beta = 0.0167$ and $\delta = 1$) by using individual-based exact continuous-time Markov model GEMF as

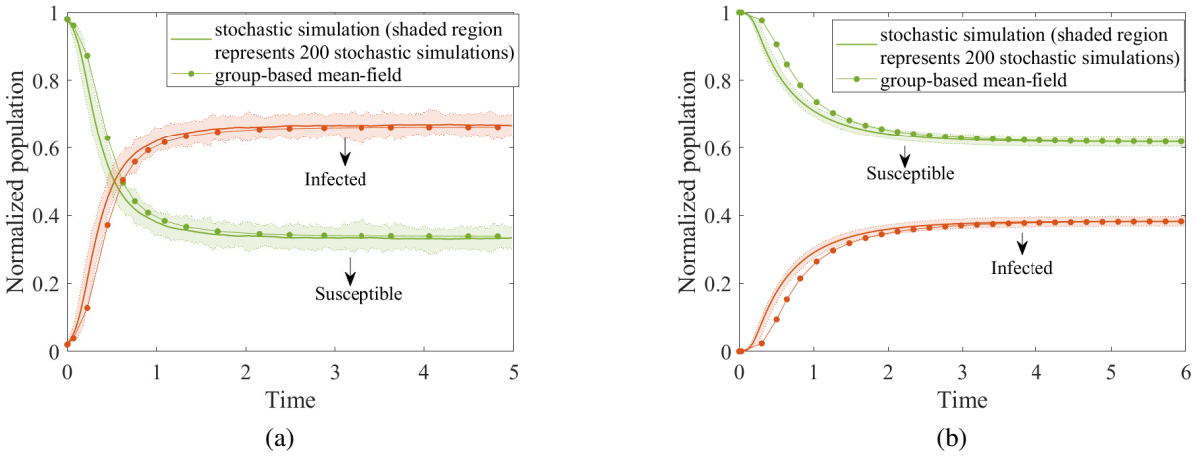
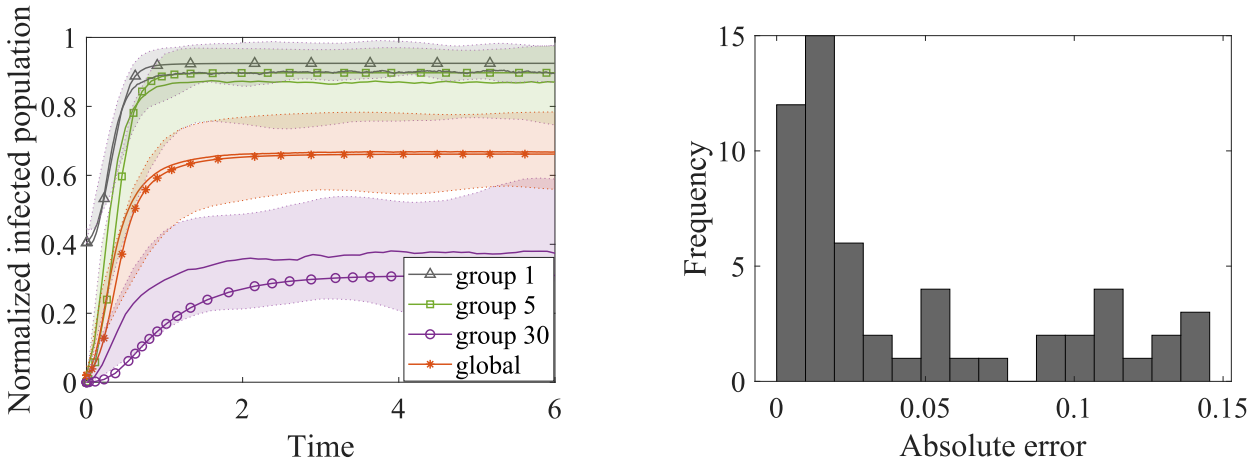


Figure 4.8: **Simulations on empirical networks for an SIS epidemic ($\beta = 0.25, \delta = 1$, node: 1-20 were infected at $t = 0$).** Solid lines represent the mean of stochastic individual-based simulation (average the dynamic of nodes of a group to compare), and shaded regions represent 200 stochastic simulations. Solid lines with markers represent the output from the intra- and inter-group mean-field model. (a) email-Eu-core network, (b) Slashdot social network, February 2009.

the benchmark. Our investigation indicates that simulation time decreases, and absolute error increases in every case with the increase of average group size, as increases of the average group size reduces the number of groups. From the investigation of absolute errors, we find that random, and k-partite approaches are more erroneous than the community and degree-based methods for the BA, and SBM network. The grouping error for the community method is close to zero in the SB and email-Eu-core network. Random grouping can not perform well in terms of accuracy in any of the networks. Absolute error for the ER network is almost zero for any grouping approaches; therefore, we do not present it in Fig 4.10.

The group-based GroupGEM framework is a solution with reduced computational time. However, because of topological and moment-closure approximation, results can deviate from the exact process. Even though this is not in the scope of this research, from previous research work, the mean-field SIS model is less accurate in sparse graphs [1]. Accuracy of the mean-field models is also sensitive to the network structure. The mean-field model can follow the exact process very closely when the size of the network is very large [127]. Extensive numerical simulation of GroupGEM in different scenarios concerning different network structures, initial conditions, and group sizes can



(a) (b)

Figure 4.9: **Local dynamics at the group level.** Simulations on the email-Eu-core network [2] (SIS epidemic model: $\beta = 0.25, \delta = 1$, node: $1 - 20$ were infected at $t = 0$). a) Time dynamics of the normalized infected population of group 1, 5, and 30. Solid lines represent the mean of stochastic individual-based simulation (average the dynamic of nodes of a group to compare), and shaded regions represent 1000 stochastic simulations. Solid lines with markers represent the output from the intra- and inter-group mean-field model. b) Histogram of absolute error of the normalized infected population of all groups at $t = 6$. Here, x-axis is the absolute error of the intra- and inter-group mean-field model compares to the mean of the stochastic individual-based simulations, and y-axis is the no. of the groups or frequency.

be a valuable research topic for future analysis.

4.6 Multilayer extension of the GroupGEM

In the real world, a contact network among interacting agents can have a complex structure, where the nature of the connection between two agents can be of multiple types. For example, in the rumor-spreading network, two people can be connected via Facebook or they can be connected via Twitter. To represents these complex structures, researchers are using multilayer networks [128, 129], where each layer represents each type of connection. If a social network has three types of connections: direct connection, Facebook connection, and Twitter connection, then a three-layer network can be used to represent this network where each layer corresponds to each type of connections. In a disease-spreading network, if a disease disperses through direct contact and by air, then a two-layer network will represent the network more precisely; one layer is for

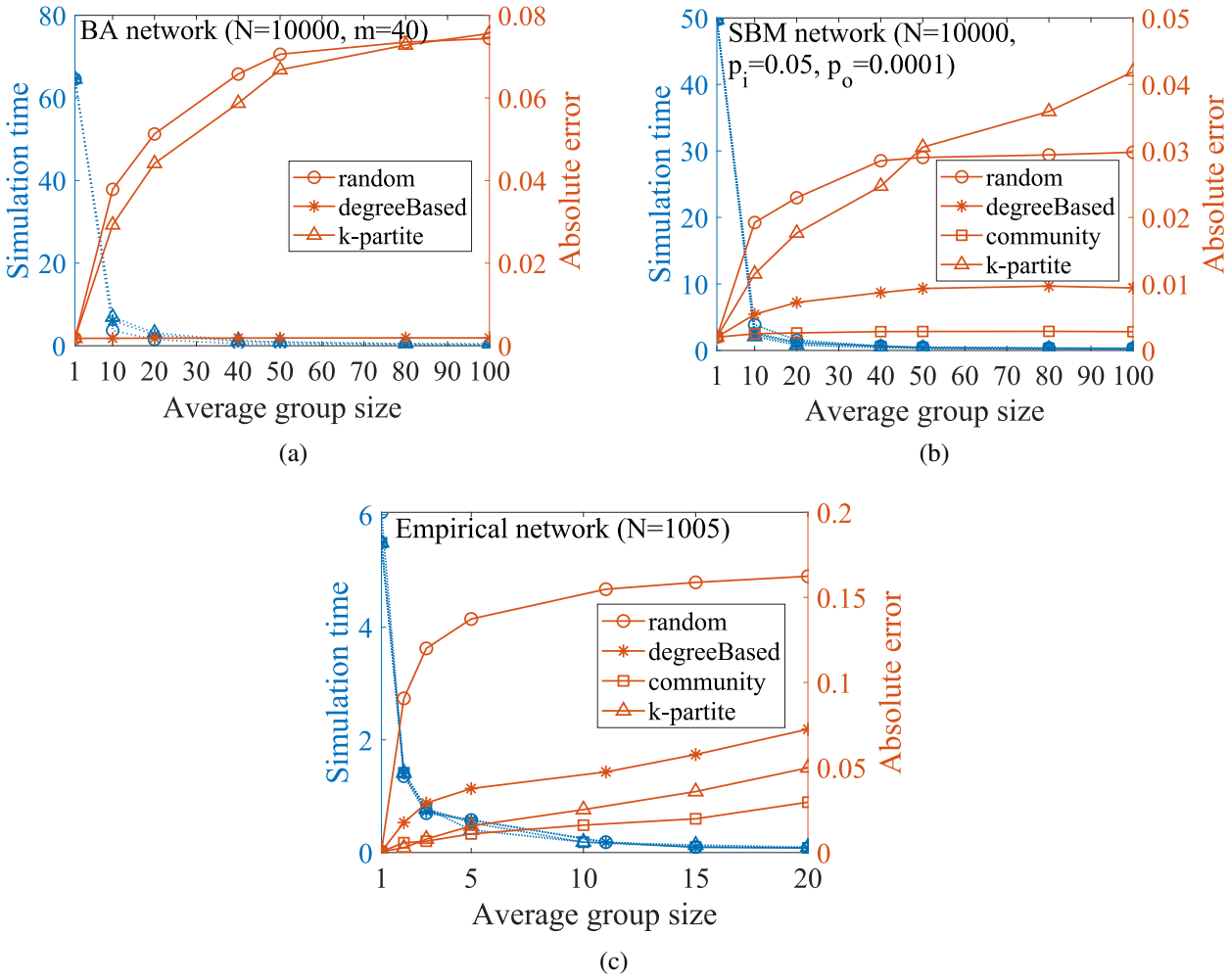


Figure 4.10: **Analysis of simulation time and absolute error for different grouping methods.** a) Barabási-Albert (BA) scale-free random network, b) stochastic-block model (SBM) network, and c) email-Eu-core empirical network.

direct contact and another layer for air transmission.

In the group-based structure, nodes and groups will be maintained in each layer; however, the connection among them will be different for different layers. An example of a group-based multilayer network, presented in Fig 4.11, has three layers. Groups are the same for each layer; however, connections are different for each layer. In particular, green lines form the link of layer-1, red lines form the link of layer-2, and purple lines form the link of layer-3.

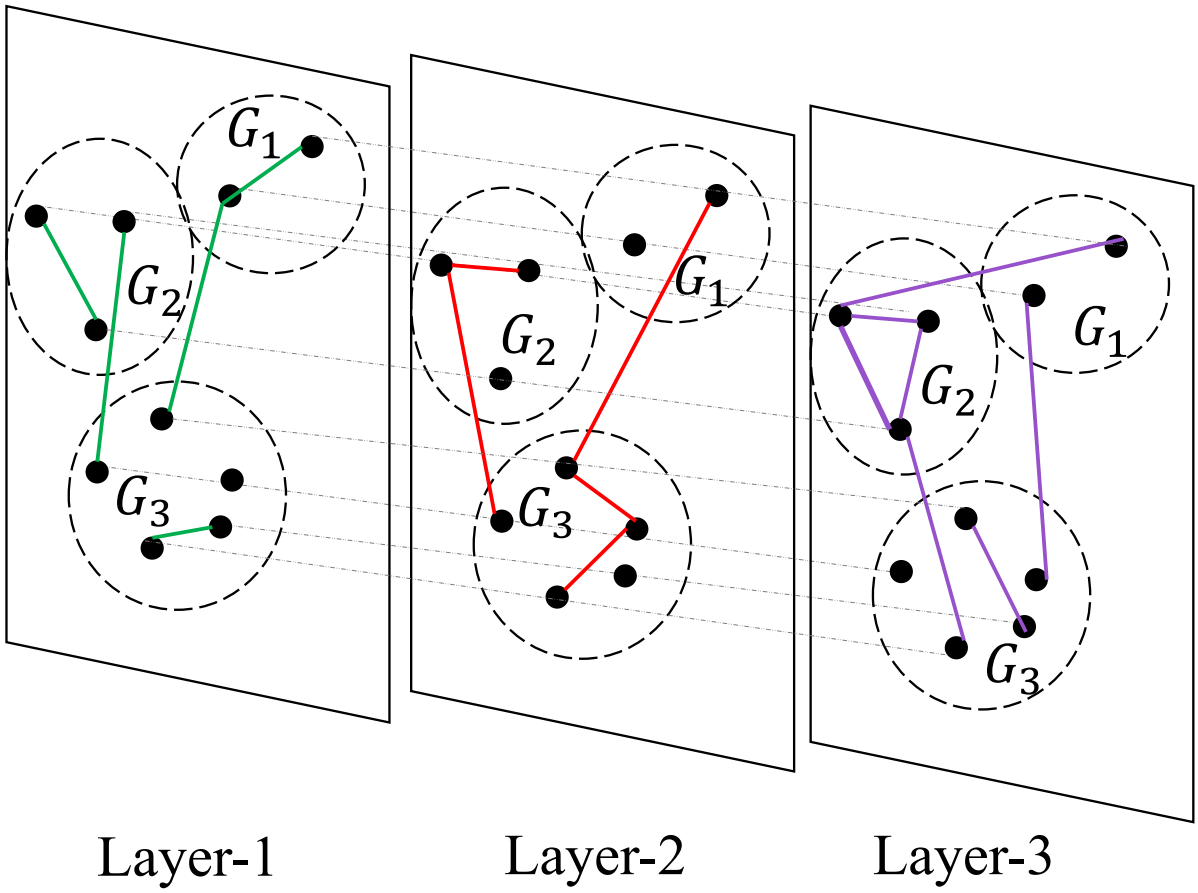


Figure 4.11: **Example of a multilayer network that has three layers.** The nodes are divided into three groups.

If the network has $\mathcal{L}$ layers, then the Eq. (4.24) can be modified as

$$\frac{d}{dt}E[g_i(t)] = \sum_{q=1}^{q_n} \Delta_{i,\delta_q}^T E[g_i(t)] + \sum_{q=1}^{q_e} \left(\sum_{l=1}^{\mathcal{L}} \Delta_{i,\beta_{ql}}^T E[h_{il}(t) g_i(t)] \right) \quad (4.30)$$

Here, the matrix for edge transition $\Delta_{i,\beta_{ql}}$ is layer-specific, and $h_{il}(t) = \sum_{j=1}^{C} A_{ij}^{gl} X_{j,r}$. The transition rate for edge transition can different for different layers. Also, the mean field equation, Eq. (4.28), can be modified for the multilayer network as

$$\frac{d}{dt}E[X_i] = \sum_{q=1}^{q_n} \Phi_{\delta_q}^T E[X_i] + \sum_{q=1}^{q_e} \left(\sum_{l=1}^{\mathcal{L}} \left(\sum_{j=1}^{C} A_{ij}^{gl} E[X_{j,r}] \right) \Phi_{\beta_{ql}}^T \right) E[X_i] \quad (4.31)$$

Only the parts for edge transition in Eq. (4.24) and (4.28) are needed to be modified for the multilayer extension, as nodal transitions are independent of the network structure. The Eq. (4.29)-C.3) can be rewritten in this similar manners as Eq. (4.31).

4.7 Future directions

In this chapter, we propose a continuous-time Markov model for the general group-based epidemic modeling (GroupGEM) framework for any compartmental model. Extensive performance analysis of the group-based approach in different types of networks with different initial conditions can be an exciting future step for this research. The group-based continuous-time Markov GroupGEM model contains a topological approximation, which is bounded by the isoperimetric inequality. However, we do not have any general guidelines for node grouping into optimal partitions. The Szemerédi's regularity lemma [123] may provide an insight to find a division where the group-based framework is expected to be accurate. A similar accuracy analysis for the SIS group-based model is described in [28]. Intuitively, if we increase the group number or decrease the group size, we will get a better division concerning accuracy. However, we cannot claim it in general.

In this research, we explore the simulation time and accuracy of the four conventional and intuitive heuristic ways to group the nodes: 1) random, 2) degree-based, 3) community, and 4) k-partite. However, there are many other ways that a practitioner can divide the nodes into groups. Therefore, choosing the optimal partition in terms of accuracy and simulation time for a specific network needs more investigation. We leave that question open for future works.

Time-varying networks can model a system more realistically but computationally expensive to handle. The researches in [130, 131] have extended the individual-based disease dispersal model [14] to consider the time-varying networks. The group-based approach can open up a possibility to reduce the computational complexity of the individual-based time-varying disease dispersal models by reducing the state space size in the Markov process.

4.8 Summary

In this chapter, we propose a general group-based epidemic modeling (GroupGEM) framework capable of representing any compartmental model in any multilayer networks. We assume a network consists of N interacting agents, which are divided into several groups. Each node can be in one of the M states. The state of a group consists of the states of its nodes. A stochastic transition of a group is caused by a stochastic transition of the state of a node. We assume each stochastic transition event is an independent Poisson process with a constant rate.

We develop a continuous-time Markov model for the group-based approach that has $\prod_{i=1}^{C} \binom{N_i+M-1}{M-1}$ possible states. This model is a multidimensional birth-death process. Possible states in the Markov chain of GroupGEM are fewer than or equal to possible states in the Markov chain of the individual-based approach, which are M^N. Therefore, GroupGEM has reduced-computational complexity and requires less simulation time with compare to the individual-based framework (GEMF).

The group-based process lies on an approximation based on the isoperimetric inequality. We further reduced the number of states by using a moment-closure approximation. The N-intertwined mean-field approximation (NIMFA) method [14, 115] and the heterogeneous mean-field method (HMF) [113, 114] are two well-known methods of the moment-closure approximation, which are two particular cases of the group-based mean-field method. The number of nonlinear differential equations for the intra- and inter-group mean-field approximation of the group-based approach is MC. Then, we present some simulation results of the GroupGEM intra- and inter-group (within- and across-group) mean-field model in synthetic networks and empirical networks. For each case, we find that simulation time reduces with the reduction of the number of groups. Finally, we provide an extension of our model for multilayer networks. This extension is important to model the dynamics of the coevolution spreading phenomena, which are seen frequently in the real world. From the coevolution of several phenomena (e.g., social distancing, propagation of disease-related information in social media, and competing viruses spreading), the estimated dynamics of the collective system can be erroneous if we only model one phenomenon and ignore others [112, 132, 133].

The GroupGEM mean-field framework lies in two approximations: topological approximation and moment-closure approximation. The topological approximation is for the underlying network, and the error for this approximation can be bounded by isoperimetric inequality. On the other hand, for the moment-closure approximation, we only know that for the $C = N$ grouping, the moment-closure approximation is the upper bound of the exact process. However, we do not have exact knowledge about the error bound for the moment-closure approximation. Accuracy of the mean-field model has been explained in [14], [1](section V).

The group-based approach allows us to scale the network and reduce computational time. It is possible to obtain the disease dynamics of an epidemic model in a large complex network by using GroupGEM when aggregated dynamics of groups of nodes are the focus of interest.

Chapter 5

Conclusion

5.1 Dissertation summary

This dissertation studies the modeling and analysis of epidemic processes over networks from data. Not only do we propose novel algorithms to infer realistic networks from limited data, but also we introduce a new computationally efficient tool to study and simulate stochastic epidemic processes over large networks.

One important element of an epidemic network model is the estimation of the network structure from data. This dissertation provides three approaches to estimate network structure from data: (1) estimates of a movement network from inventory and sales data (Chapter 2), (2) selection of the best suitable network from the incidence data (Appendix A), and (3) development of an age-specific contact network from demographic data and Google community mobility reports (Chapter 3). In the first approach, we solve a convex optimization problem by using the maximum entropy method. Then, we propose a novel algorithm to develop a higher resolution network from the lower resolution movement probabilities. The network analysis algorithms in the generated network find evidence of small-world phenomena in the US swine industry. This research also indicates what types of additional data the USDA-NASS department can collect to improve the movement estimation while retaining the data anonymity. In the second approach, we propose several realistic network models and select the one best supported from the incidence

data by developing the approximate Bayesian computation based on a sequential Monte Carlo sampling (ABC-SMC) method for network models. In the third approach, we develop an age-specific, individual-based contact network using a configuration network model approach from demographic data and Google mobility data. The generated contact network represents heterogeneous social mixing at the individual level.

Network epidemic models require disease transmission parameters, which are disease and location-specific. Chapter 3 provides a guideline to estimate unknown epidemic model parameters from incidence data using the ABC-SMC method for network models. We use this ABC-SMC scheme in our developed individual-based disease management framework to uncover the disease parameters of the COVID-19 transmission. Our framework is realistic and flexible enough to add more parameters or variables depending on the data availability. We use this disease management framework to understand the costs and benefits of contact tracing in COVID-19 transmission, specifically for the reopening phase in a small college town. Our investigation indicates that a sufficient amount of contact tracing can reduce the impact of COVID-19 dispersal in that reopening process. At first, the number of quarantined susceptible people increases with the percentage of traced contacts; however, after a certain amount of traced contacts (for example, $t\%$), the number of quarantined susceptible people starts to decrease with the increase in the percentage of traced contacts. Therefore, it is cost-effective to trace more than $t\%$ contacts of a confirmed case. This research investigates the optimum traced percentage for different movement levels.

Finally, we develop a general group-based epidemic modeling (GroupGEM) framework to reduce the computational time of the individual-based framework in Chapter 4. Our research proposes three levels of approximation to the original individual-based continuous-time Markov model that finds equations of varying lower complexity compared to that of the exact Markov equations that involve M^N states (where M is the number of compartments and N is the number of nodes):

1. Purely topological reduction can be obtained by partitioning nodes into groups. This finds exact Markov equations that provide fully joint probability distribution within (intra) and across (inter) groups. Partitioning the population into C groups reduces the state-space from

M^N states to $\prod_{i=1}^{C} \binom{N_i+M-1}{M-1}$ states, which is already polynomial in N for a constant number of groups and quasi-polynomial in N for a logarithmic number of groups, i.e., $C = O(\log N)$,

2. Inter-group mean-field approximation, where we assume states of each group are uncorrelated with other groups. This will give us the joint probability distribution of states within groups. This second level of approximation reduces the state space size to $\sum_{i=1}^{C} \binom{N_i+M-1}{M-1}$, which is polynomial in N even with a logarithmic or linear number of groups,

3. Intra- and inter-group mean-field approximation, which gives marginal state probabilities of each group yielding only $C(M-1)$ equations, which is independent of the network size N for constant C and that linearly grows with a number of groups.

The simulation results find that computational time reduces with the increase in average group size. The computational time vs. average group size plots have a knee point; computational time reduces with high speed until this knee point (Figure 4.10). A curve visibly bends in a knee point from high slope to low slope, or vice versa [134]. Our research also finds that a group-based approach can reduce the simulation time without compromising the accuracy for specific selections of groups.

5.2 Future research directions

The immediate future step of the group-based continuous-time Markov GroupGEM framework is extensive performance analysis of different types of networks with different initial conditions. The GroupGEM model contains a topological approximation, which is bounded by the isoperimetric inequality. However, we do not have any general guidelines for node grouping into optimal partitions. The Szemerédis regularity lemma [123] may provide insight into finding a division where the group-based framework is expected to be accurate. A similar accuracy analysis for the SIS group-based model is described in [28]. Intuitively, if we increase the group number or decrease the group size, we will get a better division concerning accuracy. However, we cannot claim it in general. This research explores the simulation time and accuracy of the four conventional and intuitive heuristic ways to group the nodes: 1) random, 2) degree-based, 3) community, and 4) k-partite. However, there are many other ways that a practitioner can divide the nodes into groups.

Therefore, choosing the optimal partition in terms of accuracy and simulation time for a specific network needs more investigation.

Concerning the estimation of the swine movement network, the accuracy of the estimation will increase with the available data; however, increasing data availability while maintaining data anonymity is challenging when designing a survey. Future work can focus on designing surveys to collect additional data improving estimations without hampering the data anonymity. Another use of the generated network in Chapter 2 could be the investigation of the stochastic pathogen dispersal processes [1, 11, 85, 86]. This study can help us understand the underlying mechanisms and threshold conditions of epidemic processes for various infectious livestock diseases.

Our age-specific disease management framework for COVID-19 in Chapter 3 leaves the door open to investigate the best vaccination strategy. Our framework is flexible enough to incorporate different vaccination schemes such as random mass vaccination, age-structured vaccination, and targeted vaccination Keeling and Rohani [7]. We use the disease management framework in a specific geographic region. Another interesting future application can be extending this framework to different geographic regions with different population densities to determine a rule-of-thumb for an optimum traced percentage that is valid across various locations.

In summary, modeling and analysis of epidemic processes over large networks from limited data is a promising research track, with many challenges and possibilities both in application and theory.

Bibliography

[1] Faryad Darabi Sahneh, Caterina Scoglio, and Piet Van Mieghem. Generalized epidemic mean-field model for spreading processes over multilayer complex networks. *IEEE/ACM Transactions on Networking (TON)*, 21(5):1609–1620, 2013.

[2] Jure Leskovec and Andrej Krevl. SNAP Datasets: Stanford large network dataset collection. http://snap.stanford.edu/data, June 2014.

[3] West Nile virus, Centers for Disease Control and Prevention. URL https://www.cdc.gov/westnile/index.html. Accessed: 2017-09-30.

[4] National Centers for Environmental Information. URL https://www.ncdc.noaa.gov/. Accessed: 2017-05-13.

[5] JF Clements, TS Schulenberg, MJ Iliff, D Roberson, TA Fredericks, BL Sullivan, and CL Wood. The ebird/clements checklist of birds of the world: v2015. *URL: http://www. birds. cornell. edu/clementschecklist/download/IOC*, 2015.

[6] Romualdo Pastor-Satorras, Claudio Castellano, Piet Van Mieghem, and Alessandro Vespignani. Epidemic processes in complex networks. *Reviews of modern physics*, 87(3):925, 2015.

[7] Matt J Keeling and Pejman Rohani. *Modeling infectious diseases in humans and animals*. Princeton University Press, 2011.

[8] Kimberly VanderWaal, Andres Perez, Montse Torremorrell, Robert M Morrison, and Meggan Craft. Role of animal movement and indirect contact among farms in transmission of porcine epidemic diarrhea virus. *Epidemics*, 2018.

[9] Pablo Valdes-Donoso, Kimberly VanderWaal, Lovell S Jarvis, Spencer R Wayne, and Andres M Perez. Using machine learning to predict swine movements within a regional program to improve control of infectious diseases in the us. *Frontiers in Veterinary Science*, 4: 2, 2017.

[10] Shankar Yadav, Olynk Widmar, J Nicole, and Hsin-Yi Weng. Modeling classical swine fever outbreak-related outcomes. *Frontiers in Veterinary Science*, 3:7, 2016.

[11] Sifat A Moon, Lee W Cohnstaedt, D Scott McVey, and Caterina M Scoglio. A spatio-temporal individual-based network framework for West Nile virus in the USA: spreading pattern of West Nile virus. *PLoS Computational Biology*, 15(3):e1006875, 2019.

[12] Sifat A Moon, Tanvir Ferdousi, Adrian Self, and Caterina M Scoglio. Estimation of swine movement network at farm level in the US from the Census of Agriculture data. *Scientific reports*, 9(1):6237, 2019.

[13] William Ogilvy Kermack and Anderson G McKendrick. A contribution to the mathematical theory of epidemics. *Proceedings of the royal society of london. Series A, Containing papers of a mathematical and physical character*, 115(772):700–721, 1927.

[14] Piet Van Mieghem, Jasmina Omic, and Robert Kooij. Virus spread in networks. *IEEE/ACM Transactions on Networking (TON)*, 17(1):1–14, 2009.

[15] Herbert W Hethcote. Three basic epidemiological models. In *Applied mathematical ecology*, pages 119–144. Springer, 1989.

[16] Mina Youssef and Caterina Scoglio. An individual-based approach to sir epidemics in contact networks. *Journal of Theoretical Biology*, 283(1):136–144, 2011.

[17] Klaus Dietz. Epidemics and rumours: A survey. *Journal of the Royal Statistical Society: Series A (General)*, 130(4):505–528, 1967.

[18] Tom Dietterich. Overfitting and undercomputing in machine learning. *ACM computing surveys (CSUR)*, 27(3):326–327, 1995.

[19] Sifat A Moon, Faryad Darabi Sahneh, and Caterina M Scoglio. Group-Based General Epidemic Modeling for Spreading Processes on Networks: GroupGEM. *IEEE Transactions on Network Science and Engineering*, 8(1):434–446, 2021. doi: 10.1109/TNSE.2020.3039494.

[20] Melissa A Sanchez and Sally M Blower. Uncertainty and sensitivity analysis of the basic reproductive rate: tuberculosis as an example. *American journal of epidemiology*, 145(12): 1127–1137, 1997.

[21] Helen J Wearing, Pejman Rohani, and Matt J Keeling. Appropriate models for the management of infectious diseases. *PLoS Med*, 2(7):e174, 2005.

[22] H Thomas Banks, Marie Davidian, John R Samuels, and Karyn L Sutton. An inverse problem statistical methodology summary. In *Mathematical and statistical estimation approaches in epidemiology*, pages 249–302. Springer, 2009.

[23] Yang Wang, Deepayan Chakrabarti, Chenxi Wang, and Christos Faloutsos. Epidemic spreading in real networks: An eigenvalue viewpoint. In *22nd International Symposium on Reliable Distributed Systems, 2003. Proceedings.*, pages 25–34. IEEE, 2003.

[24] Alain Barrat, Marc Barthelemy, and Alessandro Vespignani. *Dynamical processes on complex networks*. Cambridge university press, 2008.

[25] Lisa Sattenspiel and Klaus Dietz. A structured epidemic model incorporating geographic mobility among regions. *Mathematical biosciences*, 128(1-2):71–91, 1995.

[26] Matt J Keeling and Pejman Rohani. Estimating spatial coupling in epidemiological systems: a mechanistic approach. *Ecology Letters*, 5(1):20–29, 2002.

[27] Vittoria Colizza, Romualdo Pastor-Satorras, and Alessandro Vespignani. Reaction–diffusion processes and metapopulation models in heterogeneous networks. *Nature Physics*, 3(4): 276–282, 2007.

[28] K Devriendt and P Van Mieghem. Unified mean-field framework for susceptible-infected-susceptible epidemics on networks, based on graph partitioning and the isoperimetric inequality. *Physical Review E*, 96(5):052314, 2017.

[29] Phillip Schumm, Caterina Scoglio, and H Morgan Scott. An estimation of cattle movement parameters in the central states of the us. *Computers and Electronics in Agriculture*, 116: 191–200, 2015.

[30] Tina Toni, David Welch, Natalja Strelkowa, Andreas Ipsen, and Michael PH Stumpf. Approximate bayesian computation scheme for parameter inference and model selection in dynamical systems. *Journal of the Royal Society Interface*, 6(31):187–202, 2009.

[31] Katalin Csilléry, Michael GB Blum, Oscar E Gaggiotti, and Olivier François. Approximate bayesian computation (abc) in practice. *Trends in ecology & evolution*, 25(7):410–418, 2010.

[32] Mark A Beaumont. Approximate bayesian computation in evolution and ecology. *Annual review of ecology, evolution, and systematics*, 41:379–406, 2010.

[33] Mikael Sunnåker, Alberto Giovanni Busetto, Elina Numminen, Jukka Corander, Matthieu Foll, and Christophe Dessimoz. Approximate bayesian computation. *PLoS Computational Biology*, 9(1):e1002803, 2013.

[34] Pierre Del Moral, Arnaud Doucet, and Ajay Jasra. Sequential monte carlo samplers. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(3):411–436, 2006.

[35] Scott A Sisson, Yanan Fan, and Mark M Tanaka. Sequential monte carlo without likelihoods. *Proceedings of the National Academy of Sciences*, 104(6):1760–1765, 2007.

[36] Karl Mokross, Thomas B Ryder, Marina Corrêa Côrtes, Jared D Wolfe, and Philip C Stouffer. Decay of interspecific avian flock networks along a disturbance gradient in Amazonia. *Proceedings of the Royal Society B: Biological Sciences*, 281(1776):20132599, 2014.

[37] Lauren A White, James D Forester, and Meggan E Craft. Using contact networks to explore mechanisms of parasite transmission in wildlife. *Biological Reviews*, 92(1):389–409, 2017.

[38] Mark Newman. *Networks: An introduction*. Oxford university press, 2010.

[39] Sifat A Moon and Caterina M Scoglio. Contact tracing evaluation for COVID-19 transmission in the different movement levels of a rural college town in the USA. *Scientific reports*, 11(1):1–12, 2021.

[40] United States Department of Agriculture Economic Research Service, . URL https://www.ers.usda.gov/webdocs/publications/37373/30253_ldpm12501_researchbrief_002.pdf?v=41414. Accessed: 2018-10-09.

[41] Scott Dee, John Deen, Kurt Rossow, Carrie Weise, Roger Eliason, Satoshi Otake, Han Soo Joo, and Carlos Pijoan. Mechanical transmission of porcine reproductive and respiratory syndrome virus throughout a coordinated sequence of events during warm weather. *Canadian Journal of Veterinary Research*, 67(1):12, 2003.

[42] Andres M Perez, Peter R Davies, Christa K Goodell, Derald J Holtkamp, Enrique Mondaca-Fernández, Zvonimir Poljak, Steven J Tousignant, Pablo Valdes-Donoso, Jeffrey J Zimmerman, and Robert B Morrison. Lessons learned and knowledge gaps about the epidemiology and control of porcine reproductive and respiratory syndrome virus in north america. *Journal of the American Veterinary Medical Association*, 246(12):1304–1317, 2015.

[43] Paolo Bajardi, Alain Barrat, Lara Savini, and Vittoria Colizza. Optimizing surveillance for livestock disease spreading through animal movements. *Journal of the Royal Society Interface*, 9(76):2814–2825, 2012.

[44] Meggan E Craft. Infectious disease transmission and contact networks in wildlife and livestock. *Phil. Trans. R. Soc. B*, 370(1669):20140107, 2015.

[45] Jagat Narain Kapur and Hiremaglur K Kesavan. Entropy optimization principles and their applications. In *Entropy and energy dissipation in water resources*, pages 3–20. Springer, 1992.

[46] Brian D Ziebart, Andrew L Maas, J Andrew Bagnell, and Anind K Dey. Maximum entropy inverse reinforcement learning. In *AAAI*, volume 8, pages 1433–1438. Chicago, IL, USA, 2008.

[47] Steven J Phillips, Robert P Anderson, and Robert E Schapire. Maximum entropy modeling of species geographic distributions. *Ecological Modelling*, 190(3-4):231–259, 2006.

[48] Edwin T Jaynes. Information theory and statistical mechanics. *Physical Review*, 106(4): 620, 1957.

[49] Alain Barrat, Marc Barthelemy, Romualdo Pastor-Satorras, and Alessandro Vespignani. The architecture of complex weighted networks. *Proceedings of the National Academy of Sciences*, 101(11):3747–3752, 2004.

[50] Albert-Laszlo Barabasi and Zoltan N Oltvai. Network biology: understanding the cell's functional organization. *Nature Reviews Genetics*, 5(2):101, 2004.

[51] Ronny P Bartsch, Kang KL Liu, Amir Bashan, and Plamen Ch Ivanov. Network physiology: how organ systems dynamically interact. *PloS One*, 10(11):e0142143, 2015.

[52] Plamen Ch Ivanov and Ronny P Bartsch. Network physiology: mapping interactions between networks of physiologic networks. In *Networks of Networks: the last Frontier of Complexity*, pages 203–222. Springer, 2014.

[53] Amir Bashan, Ronny P Bartsch, Jan W Kantelhardt, Shlomo Havlin, and Plamen Ch Ivanov. Network physiology reveals relations between network topology and physiological function. *Nature Communications*, 3:702, 2012.

[54] Ronny P Bartsch, Aicko Y Schumann, Jan W Kantelhardt, Thomas Penzel, and Plamen Ch Ivanov. Phase transitions in physiologic coupling. *Proceedings of the National Academy of Sciences*, 109(26):10181–10186, 2012.

[55] Kyuyoung Lee, Dale Polson, Erin Lowe, Rodger Main, Derald Holtkamp, and Beatriz Martínez-López. Unraveling the contact patterns and network structure of pig shipments

in the united states and its association with porcine reproductive and respiratory syndrome virus (prrsv) outbreaks. *Preventive Veterinary Medicine*, 138:113–123, 2017.

[56] KK Thakur, CW Revie, D Hurnik, Z Poljak, and J Sanchez. Analysis of swine movement in four c anadian regions: Network structure and implications for disease spread. *Transboundary and Emerging Diseases*, 63(1):e14–e26, 2016.

[57] Fabrizio Natale, Armando Giovannini, Lara Savini, Diana Palma, Luigi Possenti, Gianluca Fiore, and Paolo Calistri. Network analysis of italian cattle trade patterns and evaluation of risks for potential disease spread. *Preventive Veterinary Medicine*, 92(4):341–350, 2009.

[58] Timothy C Germann, Kai Kadau, Ira M Longini, and Catherine A Macken. Mitigation strategies for pandemic influenza in the united states. *Proceedings of the National Academy of Sciences*, 103(15):5935–5940, 2006.

[59] Matt J Keeling, Mark EJ Woolhouse, Darren J Shaw, Louise Matthews, Margo Chase-Topping, Dan T Haydon, Stephen J Cornell, Jens Kappey, John Wilesmith, and Bryan T Grenfell. Dynamics of the 2001 uk foot and mouth epidemic: stochastic dispersal in a heterogeneous landscape. *Science*, 294(5543):813–817, 2001.

[60] United States Department of Agriculture National Agricultural Statistics Service, . URL http://usda.mannlib.cornell.edu/usda/current/hogview/hogview-10-29-2015.pdf. Accessed: 2018-04-30.

[61] United States Department of Agriculture, . URL http://usda.mannlib.cornell.edu/usda/current/hogview/hogview-10-29-2015.pdf. Accessed: 2018-04-30.

[62] Pork checkoff. URL https://www.pork.org/facts/stats/structure-and-productivity/americas-top-100-pig-counties/. Accessed: 2018-04-30.

[63] Diana María Herrera-Ibatá, Beatriz Martínez-López, Darla Quijada, Kenneth Burton, and Lina Mur. Quantitative approach for the risk assessment of african swine fever and classi-

cal swine fever introduction into the united states through legal imports of pigs and swine products. *PloS one*, 12(8):e0182850, 2017.

[64] Nailong Wu. *The maximum entropy method*, volume 32. Springer Science & Business Media, 2012.

[65] John Harte. *Maximum entropy and ecology: a theory of abundance, distribution, and energetics*. OUP Oxford, 2011.

[66] Alaa M El-Halees. Arabic text classification using maximum entropy. *IUG Journal of Natural Studies*, 15(1), 2015.

[67] Wanting Xiong, Luca Faes, and Plamen Ch Ivanov. Entropy measures, entropy estimators, and their performance in quantifying complex dynamics: Effects of artifacts, nonstationarity, and long-range correlations. *Physical Review E*, 95(6):062114, 2017.

[68] John Harte and Erica A Newman. Maximum information entropy: a foundation for ecological theory. *Trends in Ecology & Evolution*, 29(7):384–389, 2014.

[69] Albert-László Barabási et al. *Network science*. Cambridge university press, 2016.

[70] Alain Barrat, Marc Barthelemy, and Alessandro Vespignani. The architecture of complex weighted networks: Measurements and models. In *Large Scale Structure And Dynamics Of Complex Networks: From Information Technology to Finance and Natural Science*, pages 67–92. World Scientific, 2007.

[71] Mark EJ Newman. Scientific collaboration networks. ii. shortest paths, weighted networks, and centrality. *Physical Review E*, 64(1):016132, 2001.

[72] Ulrik Brandes. A faster algorithm for betweenness centrality. *Journal of Mathematical Sociology*, 25(2):163–177, 2001.

[73] Geoffrey S Canright and Kenth Engø-Monsen. Spreading on networks: a topographic view. *Complexus*, 3(1-3):131–146, 2006.

[74] Johannes J Bisschop and Robert Entriken. *AIMMS: The modeling system*. Paragon Decision Technology BV, 1993.

[75] Mathieu Bastian, Sebastien Heymann, Mathieu Jacomy, et al. Gephi: an open source software for exploring and manipulating networks. *ICWSM*, 8(2009):361–362, 2009.

[76] Thomas MJ Fruchterman and Edward M Reingold. Graph drawing by force-directed placement. *Software: Practice and experience*, 21(11):1129–1164, 1991.

[77] United States Department of Agriculture Animal and Plant Health Inspection Service, . URL https://www.aphis.usda.gov/animal_health/nahms/smallscale/downloads/Small_Producer.pdf. Accessed: 2018-10-12.

[78] J Giamalva. Pork and swine. industry and trade summary. *United States International Trade Commission*, 2014.

[79] Nigel Key and William McBride. The changing economics of us hog production. 2007.

[80] William McBride and Nigel Key. Characteristics and production costs of us hog farms, 2004. 2007.

[81] Istvan Z Kiss, Darren M Green, and Rowland R Kao. Infectious disease control using contact tracing in random and scale-free networks. *Journal of The Royal Society Interface*, 3(6):55–62, 2006.

[82] Anand Nair and José M Vidal. Supply network topology and robustness against disruptions–an investigation using multi-agent model. *International Journal of Production Research*, 49 (5):1391–1404, 2011.

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

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

[85] Narges Montazeri Shahtori, Tanvir Ferdousi, Caterina Scoglio, and Faryad Darabi Sahneh. Quantifying the impact of early-stage contact tracing on controlling ebola diffusion. *Mathematical Biosciences & Engineering*, 15(5):1165–1180, 2018.

[86] Tanvir Ferdousi, Lee W Cohnstaedt, DS McVey, and Caterina M Scoglio. Understanding the survival of zika virus in a vector interconnected sexual contact network. *bioRxiv*, page 518613, 2019.

[87] Tanvir Ferdousi, Sifat Afroj Moon, Adrian Self, and Caterina Scoglio. Generation of swine movement network and analysis of efficient mitigation strategies for African swine fever virus. *PloS One*, 14(12), 2019.

[88] Catrin Sohrabi, Zaid Alsafi, Niamh O'Neill, Mehdi Khan, Ahmed Kerwan, Ahmed Al-Jabir, Christos Iosifidis, and Riaz Agha. World health organization declares global emergency: A review of the 2019 novel coronavirus (covid-19). *International Journal of Surgery*, 2020.

[89] World health organization. https://covid19.who.int/. Accessed: 2020-05-31.

[90] U.S. Census Bureau (2018). Selected social characteristics in the united states. https://data.census.gov/cedsci/. Accessed: 2020-03-30.

[91] Google covid-19 community mobility reports. https://www.google.com/covid19/mobility/. Accessed: 2020-05-31.

[92] C Barril, À Calsina, and J Ripoll. A practical approach to r 0 in continuous-time ecological models. *Mathematical Methods in the Applied Sciences*, 41(18):8432–8445, 2018.

[93] Dimitri Breda, Francesco Florian, Jordi Ripoll, and Rossana Vermiglio. Efficient numerical computation of the basic reproduction number for structured populations. *Journal of Computational and Applied Mathematics*, 384:113165, 2021. doi: https://doi.org/10.1016/j.cam.2020.113165.

[94] NW Furukawa, JT Brooks, and J Sobel. Evidence supporting transmission of severe acute

respiratory syndrome coronavirus 2 while presymptomatic or asymptomatic. *Emerging infectious diseases*, 26(7), 2020.

[95] Centers for disease control and prevention. https://www.cdc.gov/coronavirus/2019-ncov/index.html, . Accessed: 2020-05-31.

[96] Joel Hellewell, Sam Abbott, Amy Gimma, Nikos I Bosse, Christopher I Jarvis, Timothy W Russell, James D Munday, Adam J Kucharski, W John Edmunds, Fiona Sun, et al. Feasibility of controlling covid-19 outbreaks by isolation of cases and contacts. *The Lancet Global Health*, 2020.

[97] Mark Newman. *Networks*. Oxford university press, 2018.

[98] Sara Y Del Valle, James M Hyman, Herbert W Hethcote, and Stephen G Eubank. Mixing patterns between age groups in social networks. *Social Networks*, 29(4):539–554, 2007.

[99] Jacco Wallinga, Peter Teunis, and Mirjam Kretzschmar. Using data on social contacts to estimate age-specific transmission parameters for respiratory-spread infectious agents. *American journal of epidemiology*, 164(10):936–944, 2006.

[100] Jonathan M Read, Ken TD Eames, and W John Edmunds. Dynamic social networks and the implications for the spread of infectious disease. *Journal of The Royal Society Interface*, 5 (26):1001–1007, 2008.

[101] Riley county health department; local health order no.3 issued march 27,2020 "stay at home". https://www.rileycountyks.gov/DocumentCenter/View/18553/03-27-2020---STAY-AT-HOME-ORDER-FROM-LOCAL-HEALTH-OFFICER-pdf. Accessed: 2020-05-31.

[102] Tobias S Brett and Pejman Rohani. Transmission dynamics reveal the impracticality of covid-19 herd immunity strategies. *Proceedings of the National Academy of Sciences*, 117 (41):25897–25903, 2020.

[103] Stephen A Lauer, Kyra H Grantz, Qifang Bi, Forrest K Jones, Qulu Zheng, Hannah R Meredith, Andrew S Azman, Nicholas G Reich, and Justin Lessler. The incubation period of coronavirus disease 2019 (covid-19) from publicly reported confirmed cases: estimation and application. *Annals of internal medicine*, 2020.

[104] Covid-19 pandemic planning scenarios, centers for disease control and prevention. https://www.cdc.gov/coronavirus/2019-ncov/hcp/planning-scenarios.html, . Accessed: 2020-11-20.

[105] Faryad Darabi Sahneh, Aram Vajdi, Heman Shakeri, Futing Fan, and Caterina Scoglio. Gemfsim: A stochastic simulator for the generalized epidemic modeling framework. *Journal of computational science*, 22:36–44, 2017.

[106] Blake Gumprecht. College towns in the united states: Table. *The American College Town*, page 1, 2008.

[107] Alessandro Vespignani. Modelling dynamical processes in complex socio-technical systems. *Nature physics*, 8(1):32, 2012.

[108] Erik Volz. SIR dynamics in random networks with heterogeneous connectivity. *Journal of mathematical biology*, 56(3):293–310, 2008.

[109] Joel C Miller, Anja C Slim, and Erik M Volz. Edge-based compartmental modelling for infectious disease spread. *Journal of the Royal Society Interface*, 9(70):890–906, 2012.

[110] Jennifer Lindquist, Junling Ma, P Van den Driessche, and Frederick H Willeboordse. Effective degree network disease models. *Journal of mathematical biology*, 62(2):143–164, 2011.

[111] Brian Karrer and Mark EJ Newman. Message passing approach for general epidemic models. *Physical Review E*, 82(1):016101, 2010.

[112] Wei Wang, Quan-Hui Liu, Junhao Liang, Yanqing Hu, and Tao Zhou. Coevolution spreading in complex networks. *Physics Reports*, 2019.

[113] Romualdo Pastor-Satorras and Alessandro Vespignani. Epidemic spreading in scale-free networks. *Physical review letters*, 86(14):3200, 2001.

[114] Marián Boguná and Romualdo Pastor-Satorras. Epidemic spreading in correlated complex networks. *Physical Review E*, 66(4):047104, 2002.

[115] Deepayan Chakrabarti, Yang Wang, Chenxi Wang, Jurij Leskovec, and Christos Faloutsos. Epidemic thresholds in real networks. *ACM Transactions on Information and System Security (TISSEC)*, 10(4):1, 2008.

[116] Wei Wang, Ming Tang, H Eugene Stanley, and Lidia A Braunstein. Unification of theoretical approaches for epidemic spreading on complex networks. *Reports on Progress in Physics*, 80(3):036603, 2017.

[117] NG Van Kampen. Stochastic processes in chemistry and physics. *Chaos*, 1981.

[118] William Feller. *An introduction to probability theory and its applications*, volume 1. John Wiley & Sons, 2008.

[119] Jovan M Nahman and JM Nahman. Dependability of engineering systems. *ELEKTROPRIVREDA*, 2002(1):102–102, 2002.

[120] Amy N Langville and William J Stewart. The kronecker product and stochastic automata networks. *Journal of computational and applied mathematics*, 167(2):429–447, 2004.

[121] Fan Chung. Discrete isoperimetric inequalities. *Surveys in differential geometry*, 9(1):53–82, 2004.

[122] Karel Devriendt and Piet Van Mieghem. Tighter spectral bounds for the cut size, based on laplacian eigenvectors. *Linear Algebra and its Applications*, 2019.

[123] Reinhard Diestel. Graph theory, volume 173 of. *Graduate texts in mathematics*, page 7, 2012.

[124] Sheldon M Ross, John J Kelly, Roger J Sullivan, William James Perry, Donald Mercer, Ruth M Davis, Thomas Dell Washburn, Earl V Sager, Joseph B Boyce, and Vincent L Bristow. *Stochastic processes*, volume 2. Wiley New York, 1996.

[125] Piet Van Mieghem. *Performance analysis of complex networks and systems*. Cambridge University Press, 2014.

[126] Marek Kubale. *Graph colorings*, volume 352. American Mathematical Soc., 2004.

[127] Cong Li, Ruud van de Bovenkamp, and Piet Van Mieghem. Susceptible-infected-susceptible model: A comparison of N-intertwined and heterogeneous mean-field approximations. *Physical Review E*, 86(2):026116, 2012.

[128] Mikko Kivelä, Alex Arenas, Marc Barthelemy, James P Gleeson, Yamir Moreno, and Mason A Porter. Multilayer networks. *Journal of complex networks*, 2(3):203–271, 2014.

[129] Maciej Kurant and Patrick Thiran. Layered complex networks. *Physical review letters*, 96 (13):138701, 2006.

[130] Philip E Paré, Carolyn L Beck, and Angelia Nedić. Epidemic processes over time-varying networks. *IEEE Transactions on Control of Network Systems*, 5(3):1322–1334, 2017.

[131] Masaki Ogura and Victor M Preciado. Stability of spreading processes over time-varying large-scale networks. *IEEE Transactions on Network Science and Engineering*, 3(1):44–57, 2016.

[132] Ji Liu, Philip E Paré, Angelia Nedić, Carolyn L Beck, and Tamer Başar. On a continuous-time multi-group bi-virus model with human awareness. In *2017 IEEE 56th Annual Conference on Decision and Control (CDC)*, pages 4124–4129. IEEE, 2017.

[133] Masaki Ogura and Victor M Preciado. Epidemic processes over adaptive state-dependent networks. *Physical Review E*, 93(6):062316, 2016.

[134] Ville Satopaa, Jeannie Albrecht, David Irwin, and Barath Raghavan. Finding a" kneedle" in a haystack: Detecting knee points in system behavior. In *2011 31st international conference on distributed computing systems workshops*, pages 166–171. IEEE, 2011.

[135] Alexis Burakoff. West nile virus and other nationally notifiable arboviral diseases—united states, 2016. *MMWR. Morbidity and Mortality Weekly Report*, 67, 2018.

[136] Louis D Bergsman, James M Hyman, and Carrie A Manore. A mathematical model for the spread of west nile virus in migratory and resident birds. *Math Biosci Eng*, 13(2):401–24, 2016.

[137] Nicholas Komar, Stanley Langevin, Steven Hinten, Nicole Nemeth, Eric Edwards, Danielle Hettler, Brent Davis, Richard Bowen, and Michel Bunning. Experimental infection of north american birds with the new york 1999 strain of west nile virus. *Emerging infectious diseases*, 9(3):311, 2003.

[138] Michael J Turell, David J Dohm, Michael R Sardelis, Monica L O'guinn, Theodore G Andreadis, and Jamie A Blow. An update on the potential of north american mosquitoes (diptera: Culicidae) to transmit west nile virus. *Journal of medical entomology*, 42(1):57–62, 2005.

[139] Georg Pauli, Ursula Bauerfeind, Johannes Blümel, Reinhard Burger, Christian Drosten, Albrecht Gröner, Lutz Gürtler, Margarethe Heiden, Martin Hildebrandt, Bernd Jansen, et al. West Nile virus. *Transfusion medicine and hemotherapy*, 40(4):265, 2013.

[140] Marjorie J Wonham, Tomas de Camino-Beck, and Mark A Lewis. An epidemiological model for west nile virus: invasion analysis and control applications. *Proceedings of the Royal Society of London B: Biological Sciences*, 271(1538):501–507, 2004.

[141] Rongsong Liu, Jiangping Shuai, Jianhong Wu, and Huaiping Zhu. Modeling spatial spread of west nile virus and impact of directional dispersal of birds. *Mathematical Biosciences and Engineering*, 3(1):145, 2006.

[142] Mark Lewis, Joanna Rencławowicz, and P Van den Driessche. Traveling waves and spread rates for a west nile virus model. *Bulletin of mathematical biology*, 68(1):3–23, 2006.

[143] Benoit Durand, Gilles Balança, Thierry Baldet, and Véronique Chevalier. A metapopulation model to simulate west nile virus circulation in western africa, southern europe and the mediterranean basin. *Veterinary research*, 41(3):32, 2010.

[144] Norberto Aníbal Maidana and Hyun Mo Yang. Spatial spreading of west nile virus described by traveling waves. *Journal of theoretical biology*, 258(3):403–417, 2009.

[145] Charles S Apperson, Hassan K Hassan, Bruce A Harrison, Harry M Savage, Stephen E Aspen, Ary Farajollahi, Wayne Crans, Thomas J Daniels, Richard C Falco, Mark Benedict, et al. Host feeding patterns of established and potential mosquito vectors of west nile virus in the eastern united states. *Vector-Borne and Zoonotic Diseases*, 4(1):71–82, 2004.

[146] A Marm Kilpatrick, Peter Daszak, Matthew J Jones, Peter P Marra, and Laura D Kramer. Host heterogeneity dominates west nile virus transmission. *Proceedings of the Royal Society of London B: Biological Sciences*, 273(1599):2327–2333, 2006.

[147] Harry M Savage, Deepak Aggarwal, Charles S Apperson, Charles R Katholi, Emily Gordon, Hassan K Hassan, Michael Anderson, Dawn Charnetzky, Larry McMillen, Emily A Unnasch, et al. Host choice and west nile virus infection rates in blood-fed mosquitoes, including members of the culex pipiens complex, from memphis and shelby county, tennessee, 2002–2003. *Vector-Borne and Zoonotic Diseases*, 7(3):365–386, 2007.

[148] Gabriel L Hamer, Uriel D Kitron, Tony L Goldberg, Jeffrey D Brawn, Scott R Loss, Marilyn O Ruiz, Daniel B Hayes, and Edward D Walker. Host selection by culex pipiens mosquitoes and west nile virus amplification. *The American journal of tropical medicine and hygiene*, 80(2):268–278, 2009.

[149] Gustavo Cruz-Pacheco, Lourdes Esteva, Juan Antonio Montaõ-Hirose, and Cristobal Vargas. Modelling the dynamics of west nile virus. *Bulletin of mathematical biology*, 67(6): 1157, 2005.

[150] Nicholas B DeFelice, Eliza Little, Scott R Campbell, and Jeffrey Shaman. Ensemble forecast of human west nile virus cases and mosquito infection rates. *Nature Communications*, 8, 2017.

[151] Michael R Sardelis, Michael J Turell, David J Dohm, and Monica L O'Guinn. Vector competence of selected north american culex and coquillettidia mosquitoes for west nile virus. *Emerging infectious diseases*, 7(6):1018, 2001.

[152] Caterina M Scoglio, Claudio Bosca, Mahbubul H Riad, Faryad D Sahneh, Seth C Britch, Lee W Cohnstaedt, and Kenneth J Linthicum. Biologically informed individual-based network model for rift valley fever in the us and evaluation of mitigation strategies. *PloS one*, 11(9):e0162759, 2016.

[153] Mahbubul H Riad, Caterina M Scoglio, D Scott McVey, and Lee W Cohnstaedt. An individual-level network model for a hypothetical outbreak of japanese encephalitis in the usa. *Stochastic environmental research and risk assessment*, 31(2):353–367, 2017.

[154] Peter M Rabinowitz, Deron Galusha, Sally Vegso, Jennifer Michalove, Seppo Rinne, Matthew Scotch, and Michael Kane. Comparison of human and animal surveillance data for h5n1 influenza a in egypt 2006–2011. *PloS One*, 7(9):e43851, 2012.

[155] Hamid Reza Nasrinpour, Alexander A Reimer, Marcia R Friesen, and Robert D McLeod. Data preparation for west nile virus agent-based modelling: Protocol for processing bird population estimates and incorporating arcmap in anylogic. *JMIR research protocols*, 6(7), 2017.

[156] Marcus SC Blagrove, Cyril Caminade, Elisabeth Waldmann, Elizabeth R Sutton, Maya Wardeh, and Matthew Baylis. Co-occurrence of viruses and mosquitoes at the vectors' optimal climate range: An underestimated risk to temperate regions? *PLoS neglected tropical diseases*, 11(6):e0005604, 2017.

[157] LM Rueda, KJ Patel, RC Axtell, and RE Stinner. Temperature-dependent development and

survival rates of culex quinquefasciatus and aedes aegypti (diptera: Culicidae). *Journal of medical entomology*, 27(5):892–898, 1990.

[158] William K Reisen, Ying Fang, and Vincent M Martinez. Effects of temperature on the transmission of west nile virus by culex tarsalis (diptera: Culicidae). *Journal of medical entomology*, 43(2):309–317, 2006.

[159] William K Reisen. Effect of temperature on culex tarsalis (diptera: Culicidae) from the coachella and san joaquin valleys of california. *Journal of medical entomology*, 32(5):636–645, 1995.

[160] Albert-László Barabási. Network science book. *Network Science*, 625, 2014.

[161] Irina Chis Ster and Neil M Ferguson. Transmission parameters of the 2001 foot and mouth epidemic in great britain. *PLoS One*, 2(6):e502, 2007.

[162] Samuel Soubeyrand, Leonhard Held, Michael Höhle, and Ivan Sache. Modelling the spread in space and time of an airborne plant disease. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 57(3):253–272, 2008.

[163] Sebastian Meyer, Leonhard Held, et al. Power-law models for infectious disease spread. *The Annals of Applied Statistics*, 8(3):1612–1639, 2014.

[164] Mark EJ Newman. Power laws, pareto distributions and zipf's law. *Contemporary physics*, 46(5):323–351, 2005.

[165] Tommy Tsan-Yuk Lam, Hon S Ip, Elodie Ghedin, David E Wentworth, Rebecca A Halpin, Timothy B Stockwell, David J Spiro, Robert J Dusek, James B Bortner, Jenny Hoskins, et al. Migratory flyway and geographical distance are barriers to the gene flow of influenza virus among north american birds. *Ecology letters*, 15(1):24–33, 2012.

[166] Mathieu Fourment, Aaron E Darling, and Edward C Holmes. The impact of migratory flyways on the spread of avian influenza virus in north america. *BMC evolutionary biology*, 17(1):118, 2017.

[167] Frederick C Lincoln. *Migration of birds*. Number 16. Government Printing Office, 1999.

[168] Sotirios Tsiodras, Theodoros Kelesidis, Iosif Kelesidis, Ulf Bauchinger, and Matthew E Falagas. Human infections associated with wild birds. *Journal of Infection*, 56(2):83–98, 2008.

[169] John H Rappole and Z Hubalek. Migratory birds and west nile virus. *Journal of applied microbiology*, 94(s1):47–58, 2003.

[170] Ellen Brooks-Pollock, Gareth O Roberts, and Matt J Keeling. A dynamic model of bovine tuberculosis spread and control in great britain. *Nature*, 511(7508):228, 2014.

[171] Chris P Barnes, Daniel Silk, and Michael PH Stumpf. Bayesian design strategies for synthetic biology. *Interface focus*, 1(6):895–908, 2011.

[172] Tina Toni and Michael PH Stumpf. Simulation-based model selection for dynamical systems in systems and population biology. *Bioinformatics*, 26(1):104–110, 2009.

[173] Anis Ben Abdessalem, Nikolaos Dervilis, David Wagg, and Keith Worden. Model selection and parameter estimation in structural dynamics using approximate bayesian computation. *Mechanical Systems and Signal Processing*, 99:306–325, 2018.

[174] Robert E Kass and Adrian E Raftery. Bayes factors. *Journal of the american statistical association*, 90(430):773–795, 1995.

[175] Xavier Didelot, Richard G Everitt, Adam M Johansen, Daniel J Lawson, et al. Likelihood-free estimation of model evidence. *Bayesian analysis*, 6(1):49–76, 2011.

[176] Jean-Michel Marin, Natesh S Pillai, Christian P Robert, and Judith Rousseau. Relevant statistics for bayesian model choice. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(5):833–859, 2014.

[177] Nicholas Komar. West nile virus: epidemiology and ecology in north america. *Advances in virus research*, 61:185–234, 2003.

[178] G Dauphin and S Zientara. West nile virus: recent trends in diagnosis and vaccine development. *Vaccine*, 25(30):5563–5576, 2007.

[179] Glen D Johnson, Millicent Eidson, Kathryn Schmit, April Ellis, and Martin Kulldorff. Geographic prediction of human onset of west nile virus using dead crow clusters: an evaluation of year 2002 data in new york state. *American Journal of Epidemiology*, 163(2):171–180, 2005.

[180] Christopher C Mundt, Kathryn E Sackett, LaRae D Wallace, Christina Cowger, and Joseph P Dudley. Long-distance dispersal and accelerating waves of disease: empirical relationships. *The American Naturalist*, 173(4):456–466, 2009.

[181] Mark Kot, Mark A Lewis, and Pauline van den Driessche. Dispersal data and the spread of invading organisms. *Ecology*, 77(7):2027–2042, 1996.

[182] Kieran J Sharkey. Deterministic epidemiological models at the individual level. *Journal of Mathematical Biology*, 57(3):311–331, 2008.

Appendix A

Parameter estimation and model selection of a spatiotemporal individual-based network framework for West Nile virus by using ABC-SMC method[1]

A.1 Spatiotemporal dynamics of West Nile virus

West Nile disease (WND) is a vector-borne zoonotic disease. This virus is the most common cause of arboviral disease in the United States [135]. From 1999 to 2017, more than 48 thousand WNV disease cases were reported to the Centers for Disease Control and Prevention (CDC), and more than two thousands of these reported cases resulted in death [3]. The underlying pattern of the West Nile virus (WNV) geographic spread across the United States is not completely clear, which is a necessary step for continental or state level mitigation strategies to reduce WNV transmission. WNV is maintained in an enzootic transmission cycle between competent mosquitoes and birds. Although many bird species may be infected with WNV, the American robin is considered an

[1]Parts of this appendix are extracted and adapted from our published article [11], Copyright ©2019, PLOS Computational Biology.

important amplifier of WNV and maybe a driver of geographic spread because WNV-infected American robins have low mortality and high viremia [136, 137]. Members of the *Culex* genus of mosquito are the principal vectors of this virus in the United States [138]. Humans, horses, and other mammals can be infected with WNV. However, these infections result in relatively low virus titers (viremia); therefore, the infected animals and people are considered dead-end hosts (not capable of infecting feeding mosquitoes). Therefore, they do not have any epidemiological impact on WNV transmission or geographic spread [139].

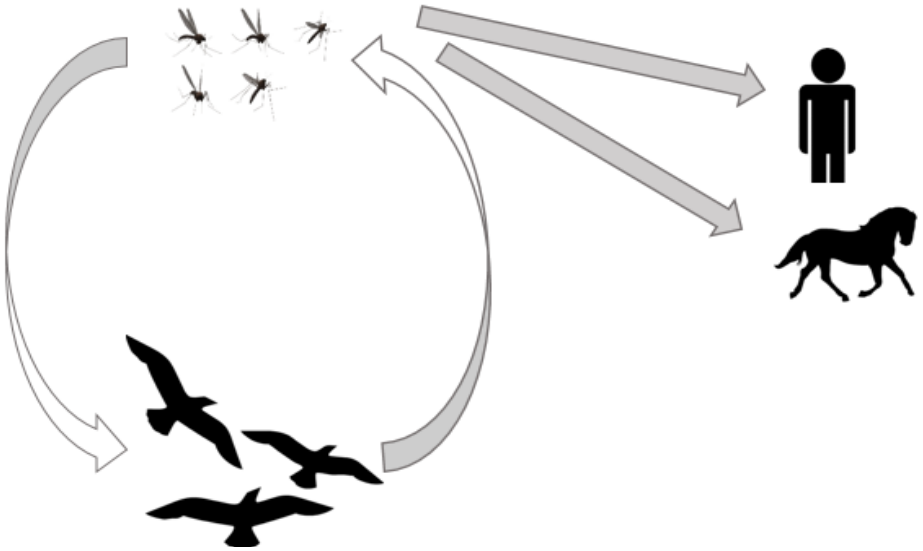


Figure A.1: **Transmission cycle of WNV.**

In the literature, several mathematical models have been developed to understand the transmission dynamics of WNV [7, 136, 140–142]. These models predict the threshold conditions for WNV spreading in different scenarios. However, most of these models do not consider the spatial dynamics of WNV. Space or geographic spread has a significant role in WNV disease dynamics and modeling of WNV spatial spreading is complex because of the interactions of multiple potential mosquito vectors, avian amplifiers, and mammalian hosts. Liu et al. [141] developed a patchy model to analyze the spatial spreading of WNV, where patches are geographical space. They assumed patches are identical, spatial dispersal of birds and mosquitoes are symmetric within patches, and movement of birds and mosquitoes are only one-dimensional. According to this investigation, long-range dispersal of infected bird populations determines the spatial spread of

WNV, not the dispersal of infected mosquito populations. Other investigators proposed a reaction-diffusion model [142], where they have spatially extended the non-spatial model of Wonham et al. [140] to mathematically estimate the spread of WNV. Here, diffusion terms in the reaction-diffusion partial differential equations represent vector mosquito and host bird population movements. They identified traveling wave solutions in their model and calculated the rate of spatial spread of infection. Durand et al. [143] developed a discrete time deterministic meta-population model in order to analyze the circulation of WNV between Southern Europe and West Africa. Another spatial model proposed by Maidana and Yang [144] used a system of partial differential reaction-diffusion equations. They also calculated the speed of disease dissemination by investigating the traveling wave solution of their model. They concluded, mosquito movements do not play an important role in disease dissemination. In addition, they included vertical transmission in their model and determined that vertical transmission is not an important factor for the spatial spread of WNV.

Most WNV spread models are mathematical deterministic compartmental models. However WNV spread is highly stochastic because of the demography and movement of hosts and vectors varies between different locations. The major weaknesses of these models are the number and complexity of the compartments required to account for the many host and vector populations. In turn, the number of compartments increases the number of unknown parameters. Approximation of these parameters in any biological system is very challenging and prone to estimation errors which can create inaccuracies in the model outputs.

We developed an individual-based heterogeneous network framework to understand WNV geographic spread. To build the network framework, we used the American Robin population density across the contiguous United States. The demographic characteristics of avian host populations and vector populations are not homogenous geographically, so we used a heterogeneous network framework. The transmission intensity of WNV depends on the abundance of WNV-infected vector mosquitoes in a given location. Mosquito population numbers fluctuate with local weather and season throughout the year, therefore we used a temperature dependent transmission rate. Although dead-end hosts cannot spread WNV to mosquitoes, we have quantified WNV case data only for humans, which we used to estimate unknown parameters.

To understand the WNV spatial distribution, we proposed distance dispersal kernels, which describes the probability of dispersal with respect to distances. In this framework, we proposed three types of distance dispersal kernels: 1) exponential, 2) power-law, and 3) power-law biased by flyway. Then we compared the three distance kernels using approximate Bayesian computation based on sequential Monte Carlo sampling (ABC-SMC) method [30–35]. After conducting an extensive simulation for 2014-2016, we observed that an adapted fat-tailed or power-law kernel, which has long-distance links in specified directions can best describe the WNV human case data [11]. We tested this network framework for the best kernel with the human case data and found that simulated results for more than 41 states of 49 states are consistent with the reported WNV cases. Our results support previous work on WNV spreading [136], which also modeled WNV spreading with migratory birds. We validate our work computationally from human incidence data. We proposed several theoretical mitigation strategies to control WNV and calculated their estimated costs. From the analysis of mitigation strategies, we suggest that potentially effective mitigation policies would include the application of mitigation control in areas with active transmission and in immediate neighboring states.

A.2 Data

The study area of this research was the contiguous United States where WNV is considered endemic. We modeled WNV case distributions for 2014-2016. We used three data sets each year to develop our model. The first dataset contained the average monthly temperatures. Mosquito vector abundance correlated with temperature. Temperature data was from the National Centers for Environmental Information [4]. The second dataset contains American Robin population data from *eBird* [5]. This is a database for bird abundance and distribution, which is formed by the Cornell Lab of Ornithology and National Audubon Society. We used total observation of American Robin in each state of the USA for each month. The robin data set was used to train the network model. The American Robin is abundant throughout the United States and is a preferred food source for many WNV-competent mosquito species [145]. Based on host feeding patterns of the *Culex* genus

of mosquitoes, robins are the most common WNV amplifying host [146–148]. Other important susceptible birds, such as American crow were not used because although they are an indicator species (high crow mortality), they are unlikely to spread virus geographically as they are mostly a residential species. In addition, as an indication of epidemic start point, we used WNV human incidence data. Many species of birds have long-distance migration during the spring and fall. Therefore the network does not focus on one long-distance migrating bird species but aggregates all species along the known flyways. To estimate model parameters we used human case data for WNV from CDC [3], which is the third dataset.

A.3 WNV epidemic model

To explore WNV long-distance spatial distribution in the USA, we used an individual-based heterogeneous network framework. In this framework, birds are on the individual level, a node represents an individual bird and connection between nodes is the possibility of virus dispersal from one infected bird to another susceptible bird by mosquito vectors. Links or connections are formed by movement of birds or movement of vectors. If there is no link between nodes then infected birds and insects are not moving virus between nodes. All virus transmission occurs by local competent vector mosquitoes. There is some evidence of bird-to-bird transmission, but it likely does not contribute to or maintain outbreaks. We split the bird population into four compartments; susceptible, exposed, infected, and recovered. Although, in the literature most mathematical models do not consider the exposed avian class when modeling WNV [140, 144, 149, 150]. Birds transmit virus to mosquitoes when a susceptible mosquito vector takes an infected blood meal, then the mosquito becomes infectious after the extrinsic incubation period (EIP), or the time needed for the virus to spreads from the mosquito mid gut to the salivary glands; usually this process takes 7 to 14 days [136, 151]. In addition, an infected bird can infect many mosquitoes simultaneously and also an infected mosquito can bite many susceptible or infected birds. Therefore, there is some delay in the system, to represent this delay we added the exposed class. We estimated exposed period from data by using the approximate Bayesian computation with sequential Monte Carlo sampling (ABC-SMC) method. After the exposed period, birds entered the infected compartment and an

infected bird transitions to recovered after 4-5 days. To simulate this model, we used generalized epidemic mean-field (GEMF) framework developed by the Network Science and Engineering (NetSE) group at Kansas State University [1]. In GEMF, each node stays in a different state and the joint state of all nodes follows a Markov process [1, 152, 153]. The node level description of this Markov process is:

$$Pr[x_i(t+\Delta t)=1|x_i(t)=0,X(t)]=\beta(T)Y_i\Delta t+o(\Delta t) \tag{A.1}$$

$$Pr[x_i(t+\Delta t)=2|x_i(t)=1,X(t)]=\lambda\Delta t+o(\Delta t) \tag{A.2}$$

$$Pr[x_i(t+\Delta t)=3|x_i(t)=2,X(t)]=\delta\Delta t+o(\Delta t) \tag{A.3}$$

Here, $X(t)$ is the joint state of all individual nodes at time t. $x_i(t)$ is a node state, $x_i(t) = C$ means node i is in C compartment at time t, $C = 0, 1, 2, 3$ corresponds to susceptible, exposed, infected, and recovered compartment. Y_i is the number of infected neighbors of node i, $\beta(T)$ is the transmission rate from one infected bird to one susceptible bird, which is a function of temperature, λ is the rate for exposed to infectious state, and finally, a node recovers from infectious state at a rate δ.

Zoonotic spillover transmission

To model disease transmission from the bird population to human population, we added a zoonotic spillover transmission compartment. We modeled occurrence of human cases as a Poisson process [150, 154]. This part of the framework can be expressed as the following equation:

$$\Delta Ih_{n_s}=Poisson(\eta Y_{n_s}) \tag{A.4}$$

In this equation, Ih_{n_s} is number of infected human cases at n sub-network in s time steps, where $s = 1, 2, 3.....$ are the discrete time steps, Y_{n_s} is infected bird population in sub-network n, and η is a scaler quantity, accounts for the contact rate and probability of pathogen transmission from bird to human. We calculated WNV spilling over to humans by using a Poisson random number

generator.

Temporal transmission rate and environmental conditions

The transmission rate for WNV is sensitive to weather data as mosquito abundance depends on the environmental conditions. Temperature, precipitation, landscape features, daylight conditions etc. are environmental conditions, which has an impact on the transmission dynamics of WNV [155]. In this research, we considered average monthly temperature data, optimal mosquito season [156], and suitable temperature range for co-occurrence of WNV and competent mosquito species. Temperature plays a very important role in the transmission dynamics of WNV because mosquito longevity and EIP are sensitive to temperature. Mosquito longevity and EIP decrease with the increase of temperature. However, there is no straightforward relationship of vectorial capacity for WNV with temperature. If incubation period decreases more than longevity, then mosquitos will be infective longer. However if longevity decreases more than incubation period, then mosquitos will not be able to transmit the virus. We used information about rainfall in this research implicitly through optical mosquito season. Optimal mosquito season of any location is estimated from monthly average temperature and rainfall data for that location [156]. In this model, we used a simple linear relation of transmission rate with temperature in a temperature window from 12°C to 32°C in the optimal mosquito season. Outside this window, transmission rate is very low. Suitable temperature for co-occurrence of WNV and *Culex pipiens* is around 12° to 27°C and for *Culex quinquefasciatus* is 20°C to 32°C [156]. Survival rate to adult stage for *Culex quinquefasciatus* is significantly high when temperature is in 20°C to 30°C [157]. For *Culex tarsalis* favorable temperature for WNV development start after 14°C [158], however larval survival reduced after 30°C temperature [159]. To compute the transmission rate of any link from node a to node b, we used temperature of the location of node b. Transmission rate for a location l is, $\beta_l(T) = \beta_\circ(T_{lm} - T_\circ)$; here, $\beta_\circ$ is the proportional constant, what we estimated by using ABC-SMC method, T_{lm} is the average temperature for month m in location l and $T_\circ$ is the threshold temperature. Threshold temperature for this model is 12°C. As the temperature is space dependent, our transmission rate also differs across the network. This individual-level heterogeneous network model gives us this

flexibility to use different transmission rate at a time for different parts of the network.

A.4 Network framework

For the spatial dynamic characteristics of WNV transmission, we built a network framework, which has 49 sub-networks one for each adjoining states of the contiguous United States plus the District of Columbia. The number of nodes in each sub-network is proportional to the size of the avian population in that state [5]. We considered the mosquito season June-October for the simulation period. Although the mosquito season is not the same for all states, mosquitoes are active from June to September in all of the states at these times [156].

The network for the avian population is (V, E). Here, V is the set of nodes, which is the union of nodes of all sub-network, $V = SN1 \cup SN2 \cup SN3 \cup ..... \cup SN49$, here SNi is a set of nodes in the sub-network i and E is the set of links among individual nodes. To build sub-networks, we used the total number of observations of American Robin for states per month in the simulation time period. $|SNi| = \max_{mj=m1:m2}(OBS^{i}_{mj}) * S_c + N_0$, here, OBS^{i}_{mj} is the total number of observations of American Robins in state i in month mj, N_0 is the error term and $N_0 \sim N(5, 2)$ for this model. $m1$ is the first month after May and $m2$ is the last month before October when the average monthly temperature is greater than T_0. S_c is the scaling constant.

In each sub-network, we assumed that nodes are connected through *Erdos-Renyi (n,p)* random network topology [83]. In this network topology, we created links randomly among nodes with a probability p. Here, n is the number of nodes in a sub-network and p is the probability to form an edge. We set the probability $p = R * log(n)/n$, here R is a constant ($R \geq 2$), as this value is more than the threshold value for the connectedness of an *Erdos-Renyi* graph [160], so nodes of a sub-network are locally connected. We will refer these networks as a local network in the subsequent sections of this appendix. To build connections among sub-networks, we considered long-distance dispersal kernels [7, 161], which describe the probability of dispersal with respect to distances. Dispersal kernels provide a simple model of dispersal to model dispersal events. For long-distance events, we used three types of kernel models; 1) Exponential, 2) power-law, and 3) power-law-flyway, which is a power-law kernel biased by flyway. The dispersal phenomenon in

this work is not conserved because of long-distance movement of migratory birds and seasonality within bird populations. Some long-distance migratory birds can disperse outside the contiguous United States or outside the network nodes, which are discrete points. The connection probability between two nodes does not represent the probability that a single dispersal event happens rather it represents the probability of contact and subsequent pathogen transmission between them. A simple caricature of the network is shown in Fig A.2. There are three sub-networks, A, B, and C. The links, which formed local networks are shown by solid lines. These links are introduced by *Erdos-Renyi (n,p)* network topology. Dashed lines are inter-links among sub-networks. These links established by using long-distance dispersal kernels.

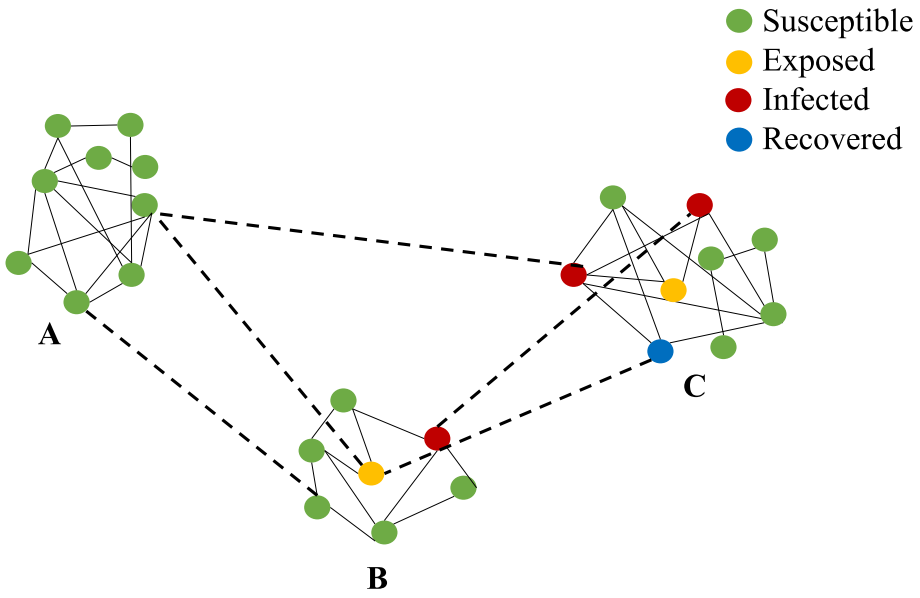


Figure A.2: **A simple caricature of the avian contact network for susceptible-exposed-infected-recovered (SEIR) epidemic model.** Here, A, B, C are three sub-networks. Solid lines represent intra-links in a sub-network and dashed lines represent inter-sub-network links.

Exponential distance kernel

In this distance kernel, connection probability among sub-networks will decrease exponentially with distance. Probability to form a link is:

$$P(d_{ij}) = K_e * \exp(-K_e * d_{ij}) \tag{A.5}$$

Here, d_{ij} is the distance between sub-network i and j, K_e is the shape parameter of exponential distribution kernel. For distance between two states, we took the distance between their centroids. The network with the exponential dispersal kernel was created as follows:

Step 1 Calculate the distance among sub-networks. d_{ij} is the distance between sub-network i and j.

Step 2 Calculate $P(d_{ij})$, this is the probability to form a link between sub-network i and j.

Step 3 Generate a random number *rand* for each pair of nodes *(a,b)*, where $a \in i$ and $b \in j$.

Step 4 If $rand < P(d_{ij})$ then an undirected link will form between node a and b.

Inter-links among sub-networks, generated by exponential distance kernel are shown in Fig A.3.

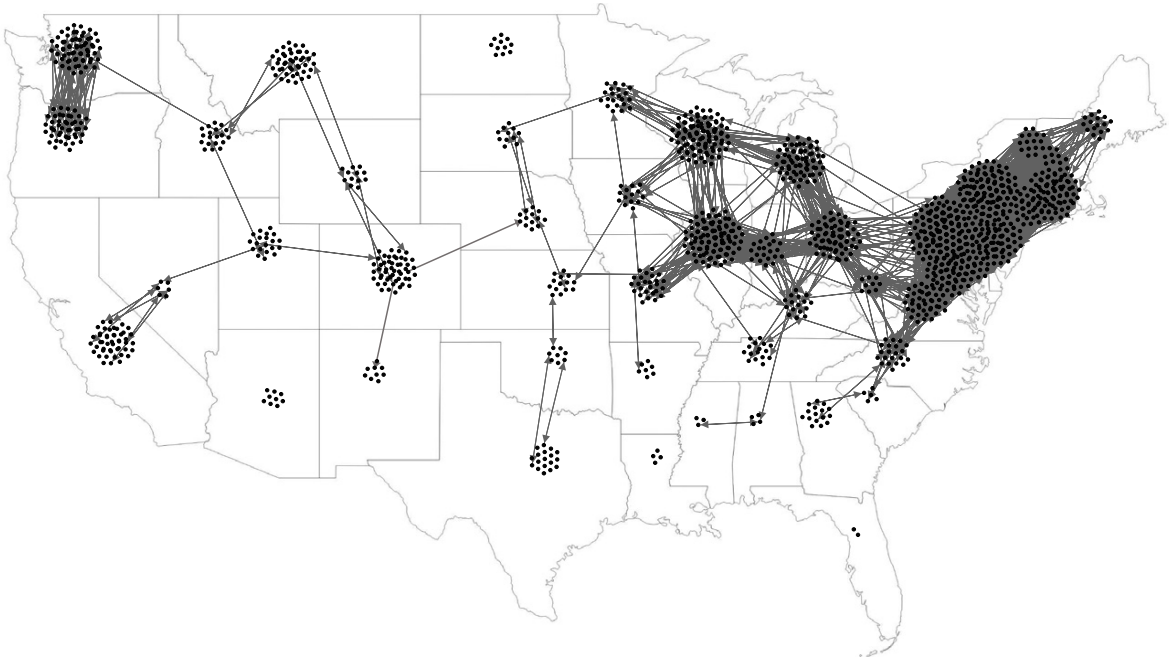


Figure A.3: **Inter-links among sub-networks for exponential distance kernel.** Links are undirected. Intra-links are not visible here. This is one realization of the stochastic networks, which is rescaled by 0.1 for better visualization.

Power-law distance kernel

Power-Law, heavy-tailed, or fat-tailed distribution allows occasional long-range transmissions of infection with frequent short-range transmissions. In this fat-tailed distance kernel, there is a

greater chance of creating links over the same long-distances compared to the exponential kernel. Power-law transmission kernel was used previously to model spatial dynamics of several infectious diseases, for example, in plant epidemiology [162], in 2001 foot-and-mouth disease epidemic [161], and also, in human diseases [163]. In power-law connections [164], the probability of connectivity among sub-networks will decrease with distance according to the following equation:

$$P(d_{ij}) = (K_{pl} - 1)/d_{min} * (d_{ij}/d_{min})^{-K_{pl}} \tag{A.6}$$

Here d_{min} is minimum distance among sub-networks and K_{pl} is the power-law parameter. The process to build this network is similar to a network for exponential kernel with the only difference being the calculation of $P(d_{ij})$. Inter-links among sub-networks for power-law distance kernel are shown in Fig A.4.

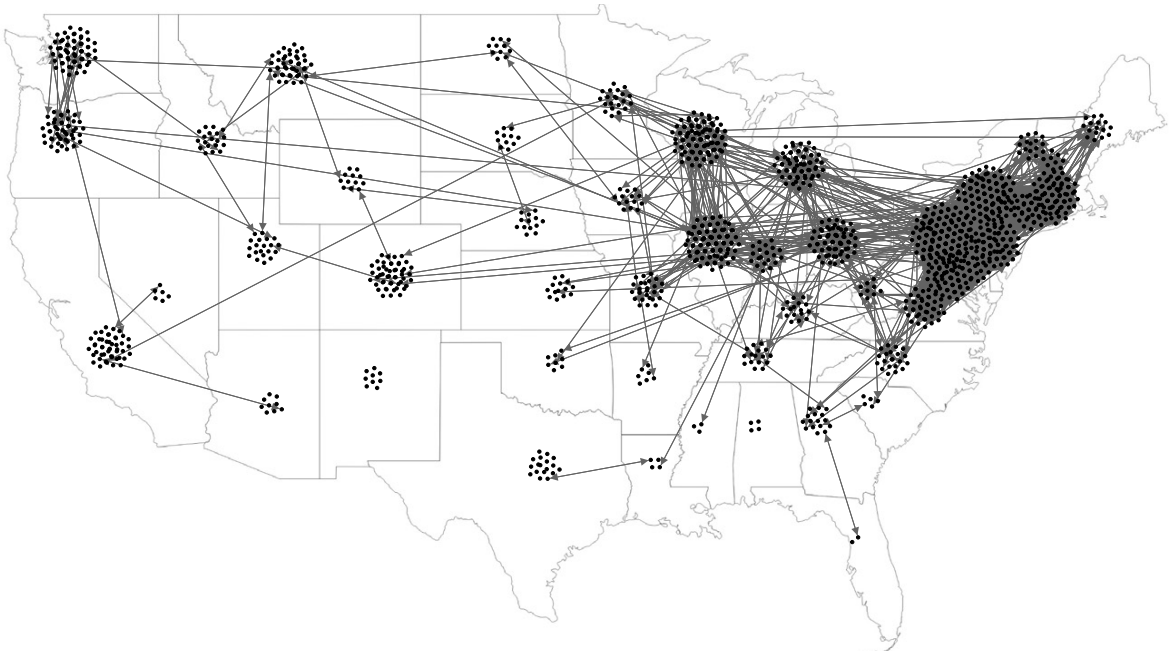


Figure A.4: **Inter-links among sub-networks for power-law distance kernel.** Links are undirected. Intra-links are not visible here. This is one realization of the stochastic networks, which is rescaled by 0.1 for better visualization.

Power-law distance kernel biased by flyway

To form this distance kernel, we included the migratory behavior of birds. Migratory birds can spread pathogens during the migration periods [165, 166]. According to the United States Fish and Wildlife Services and Flyway Councils, there are four flyways in the United States; the Atlantic flyway (AF), the Mississippi flyway (MF), the Central flyway (CF), and the Pacific flyway (PF) [167]. Although flyways overlap and the migratory patterns are very complex, these migratory routes play a vital role in the long-distance spreading of WNV [168]. To build this distance kernel, we considered two types of links among sub-networks; 1) links which are formed for residential or short-distance migratory bird movements and 2) links which are formed for long-distance migratory bird movements. For the first type of links, we used an estimated movement range of 500 km [169], these connections are unrelated to flyways. For the second type of connections, we considered two migration periods; spring migration (April - June) and late summer/fall migration (July - September) [153]; during the spring migration, we established long links from south to north and in late summer/fall migration, the reverse. To establish any long link, we picked two sub-network and establish a link if they were in the same flyway with probability $P(d_{ij})$ (Eq. A.6), these links were directional and direction was imposed with respect to migratory period. Inter-links among sub-networks for this kernel were shown in Fig A.5. The algorithm to create this network was:

Step 1 Calculate the distance among sub-networks. d_{ij} is the distance between sub-network i and j.

Step 2 Calculate $P(d_{ij})$ using Eq. A.6, this is the probability to form a link between states i and j.

Step 3 Generate a random number *rand* for each pair of nodes *(a,b)*, where $a \in i$ and $b \in j$.

Step 4 If $rand < P(d_{ij})$ and $d_{ij} < 500km$ then an undirected link will form between node a and b.

Step 5 If $rand < P(d_{ij})$ and $d_{ij} > 500km$ and states i and j are in the same flyway then an directed link will form between node a and b according to the migration period.

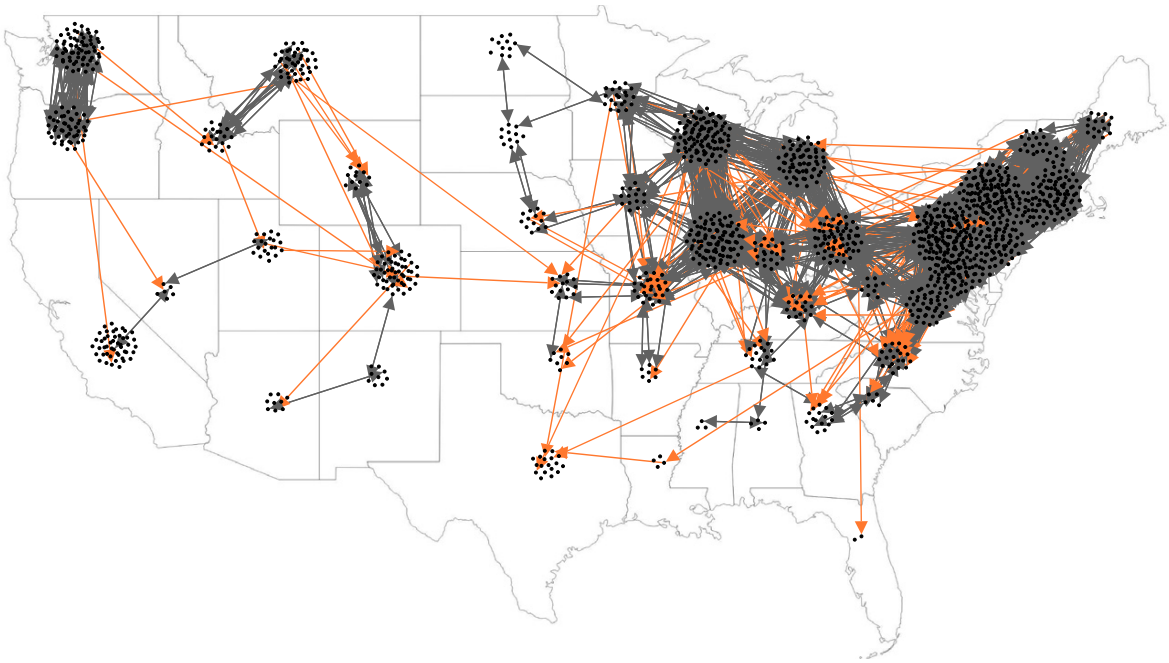


Figure A.5: **Inter-links among sub-networks for power-law distance kernel biased by flyway.** Gray links represent undirected links and orange links represent directed links (for spring migration –northbound; for late summer/fall migration –southbound). Intra-links are not visible here. This is one realization of the stochastic networks, which is rescaled by 0.1 for better visualization.

Temporal network behavior

Bird populations are not constant in any region, they change with time because of bird movement. To consider this fact, this study adds a node property, namely, *Activity*. This property can hold two values: 1 = *Active* and 0 = *Inactive*. In the entire network, only *Active* node can contribute to the spreading of the WNV. By controlling this property, we varied the size of the active node population in any sub-network with respect to the variation of the avian population in that region. The length of the simulation each year was five months (June - October). Then, each month nodes are activated randomly according to the total number of birds observed in that region in that month.

A.5 ABC-SMC for parameter estimation and model comparison

In this framework, we adopted approximate Bayesian computation based on a sequential Monte Carlo sampling (ABC-SMC) method for parameter estimation and model selection [30–35].

A.5.1 Parameter estimation

ABC-SMC is a computational method of Bayesian statistics that combines a particle filtering method with summary statistics. This method is ideal for a stochastic complex model where likelihood function is intractable or computationally expensive to evaluate. ABC estimates the posterior distribution of parameters from data. Let, θ is a parameter vector to be estimated. The goal of the ABC is to approximate the posterior distribution, $\Pi(\theta|d) \propto f(d|\theta)\Pi(\theta)$, where prior distribution of parameters $\Pi(\theta)$ are given and $f(d|\theta)$ is the likelihood of θ given the data d. This method samples parameter values from their prior distribution through subsequent SMC rounds. Intermediate distribution of the parameter is $\Pi(\theta|dist(x,d) \leq \epsilon_i); i = 1, 2, ...P$. The target posterior distribution is $\Pi(\theta|dist(x,d) \leq \epsilon_P)$. Here, x is the simulated data set, $dist$ is the distance function, ϵ is the tolerance and P is the number of SMC rounds or the number of populations, where $\epsilon_P < ..... < \epsilon_2 < \epsilon_1$ [170]. This is an adapted sequential importance sampling. In each SMC round, it uses perturbation kernel to sample a parameter set. After each simulation of the model, the model output and data are compared using some goodness-of-fit metrics. A parameter set is accepted if the distance between the model output and data is less than the tolerance level. The accepted parameter set is a particle and accepted particles form a population for that SMC round. We used two goodness-of-fit metric or distance function in this research. The first goodness-of-fit metric is squared root of the sum of squared error between observed incidence data and simulated incidence data for any proposed parameter set. The first goodness-of-fit metric for this model is:

$$dist_1(x,d) = \sqrt{\sum_{i=1}^{w}\sum_{j=1}^{s}(x(i,j) - d(i,j))^2} \tag{A.7}$$

Here, *x(i,j)* is simulated incidence model data for i week and for j location. The second goodness-of-fit metric is the absolute difference between the number of infected states from observed data and simulated data, infected state defined as a state where at least one infected individual has reported. The ABC-SMC algorithm, we adapted for this model from Toni et al. [30]. The steps for approximate Bayesian computation with sequential Monte Carlo sampling (ABC-SMC) algorithm for parameter estimation are [30–35]:

Step 1 Initialize tolerance ϵ for each SMC round, where $\epsilon_P < ..... < \epsilon_2 < \epsilon_1$. Set Population indicator, p=1.

Step 2 Particle indicator, n=1.

Step 3 Generate a particle (set of parameters),θ_p^n

(a). if p=1, sample from prior of parameters, $\pi(\theta)$;

(b). if p>1, sample the particle from previous population $\{\theta_{p-1}^n\}$ with weights W_{p-1} and then perturb the particle, $\theta^{'}$ by using perturbation kernel, PK_p to get $\theta^{''}$.

(c). if $\pi(\theta^{''}) == 0$, return to Step 3.

Step 4 Run the model R times with the new particle and compare the simulated weekly human WNV incidence with observed weekly WNV incidence using the goodness-of- fit metric, We calculated $r_p(\theta^{''}) = (1/R) * \sum_{r=1}^{R} 1(dist(x,d) < \epsilon_p)$, if $r_p(\theta^{''}) == 0$ reject the particle; go back to Step 3(a).

Step 5 Calculate the weight for the accepted particle,

(a). if p=1, $W_{n,p} = r_p(\theta^{''})$;

(b). if p>1, the weight is given by, $W_{n,p} = \frac{\pi(\theta_p^i) * r_p(\theta^{''})}{\sum_{j=1}^{N} W_{i,p-1} PK_p(\theta_{p-1}^j, \theta_p^i)}$.

Step 6 Repeat steps 3 - 5 until N= 1000 particles have been accepted.

Step 7 Normalize the weights. If $p < P$, set p= p+1, go to Step 2.

We used this algorithm separately for estimating parameters for this three distance dispersal kernel network models. As our models are an event based stochastic simulation, we simulated them 30 times with GEMF for each particle to get 30 realizations of the system. Then we take the average of these realizations. As the average over the multiple runs of a stochastic system holds more information than a single stochastic run.

A.5.2 Model comparison

In many areas, researchers deal with model selection. Bayesian theory is a comprehensive method to make inference about models from data. Approximate Bayesian computation was used in many research areas for model selection [171]. To compare among three distance kernels, this investigation used ABC-SMC model selection framework [30, 172, 173]. For given data d, the marginal posterior probability of model m is:

$$Pr(m|d) = Pr(d|m)Pr(m)/Pr(d) \tag{A.8}$$

Here, $Pr(d|m)$ is the marginal likelihood and $Pr(m)$ is the prior probability of the model. We used a uniform distribution for prior distribution of unknown parameters. For each model, we have four unknown parameters; network parameter K (K_e is the network parameter for the exponential kernel and K_{pl} is the network parameter for the both power-law kernels), constant for transmission rate β_0, transition rate from exposed to infectious state λ, and zoonotic transmission spillover rate η. In each population, we took 1000 particles. We used Bayes factor to compare a model with another model. For model m_i and m_j, Bayes factor [174] is,

$$B_{ij} = \frac{Pr(m_i|d)/Pr(m_j|d)}{Pr(m_i)/Pr(m_j)}, \tag{A.9}$$

Here, $Pr(m_i)$ is the prior and $Pr(m_i|d)$ is the marginal posterior distribution of model m_i. The Bayes factor is a summary evidence in favor of one model over another supported by the data. If B_{ij} is in range 1-3, we can conclude that summary of the evidence against mj in favor of m_i is very weak. If B_{ij} is in range 3-20, we can conclude that summary of the evidence against m_j in favor of m_i

is positive [174]. The ABC-SMC model selection algorithm is very similar to the algorithm for parameter estimation. Here, m is the model indicator, $m \in 1, 2, ..., M$, M is the number of model. In this research, we had three network models ($M = 3$) to compare.

m = 1: exponential kernel network model,

m = 2: power-law kernel network model, and

m = 3: power-law kernel influenced by flyway network model.

In each population, the model selection algorithm starts by sampling the model parameter m from the prior distribution $\Pi(m)$. Then the algorithm proposes a new set of parameters (particle) from the sets of parameters of the model m from the previous population. The Bayes factor was calculated from the final population of m.

The steps for approximate Bayesian computation with sequential Monte Carlo sampling (ABC-SMC) algorithm for model selection are [30–35]:

Step 1 Initialize tolerance for each SMC round $\epsilon_P < ..... < \epsilon_2 < \epsilon_1$. Set Population indicator, p=1.

Step 2 Particle indicator, n=1.

Step 3 Generate a particle

(a). if p=1, sample model parameter m and parameters for that model from prior, $\pi(m, \theta)$;

(b). if p>1, sample model $m^{'}$ with probability $Pr_{t-1}(m^{'})$ and then perturb by perturbation kernel PKm_p, sample the particle from previous population $\{\theta(m^{''})_{p-1}\}$ with weights W_{p-1} and then perturb the particle $\theta^{'}$ by using perturbation kernel to get $\theta^{''}$.

(c). if $\pi(m^{''}, \theta^{''}) == 0$, return to Step 3.

Step 4 Run the model $m^{''}$, R times with the new particle and compare the simulated weekly human WNV incidence with observed weekly WNV incidence using the goodness-of- fit metric, We calculated $r_p(\theta^{''}) = (1/R) * \sum_{r=1}^{R} 1(dist(x, d) < \epsilon_p)$, if $r_p(\theta^{''}) == 0$ reject the particle; go back to Step 3.

Step 5 Calculate weight for the accepted particle, set $(m_p^n, \theta_p^n) = (m^{''}, \theta^{''})$,

(a). if p=1, $W_{n,p}(m_p^n, \theta_p^n) = (1/R) * \sum_{r=1}^{R} 1(dist(x,d) < \epsilon_p)$; Here R is the number of replicate simulation run for a fixed particle.

(b). if p>1, the weight is given by, $W_{n,p}(m_p^n, \theta_p^n) = \frac{\pi(m_p^n,\theta_p^n)*(1/R)*\sum_{r=1}^{R} 1(dist(x,d)<\epsilon_p)}{\sum_{j=1}^{N} W_{i,p-1} PK_p(\theta^j_{p-1},\theta^i_p)}$.

Step 6 Repeat steps 3 - 5 until N= 1000 particles have been accepted.

Step 7 Normalize the weights for every m. If $p < P$, set p= p+1, go to Step 3.

Although ABC-SMC is an accurate statistical tool for parameter estimation and model selection, however, the results of this method are sensitive to summary statistics [175]. For our case, no summary statistics were required because we used the entire set of data and we compared the simulated and observed dataset directly by using goodness-of-fit or distance metric. A full dataset is sufficient to get the consistent result from approximate Bayesian Computation [176].

A.6 Mitigation strategies

The role of mosquito populations in WNV transmission is expressed by disease transmission rate β. This framework used different transmission rates in different parts of the network corresponding to the local mosquito abundance. Using this heterogeneous feature in the framework, we evaluated theoretical mosquito population management measures to reduce the outbreak size or transmission rates in the state level. Some states such as Kansas, do not have statewide mosquito surveillance or management, but in these theoretical scenarios, it is assumed they can develop or benefit from effective statewide mosquito management programs. The framework will simply estimate how much the mosquito abundance is reduced or maintained based on the theoretical outcomes of coordinated control. Furthermore, we realize mosquito control is generally conducted on a county or municipal level, but the human case data is only available on a state level. Therefore the recommendations are for the lowest resolution of the data, which is state level but applies to counties and municipalities as well. If vector management is increased in a sub-network, then transmission rates will be changed by, $\beta_r = \frac{\beta}{RF}$, here β_r is the reduced transmission rate and RF is the reduction factor. Then management costs will be $Cost = RF * NS_c$, here NS_c is the number of states where

control measures were applied. We considered supplemental management measures with the existing management measures. We used two types of mitigation strategies across the United States, 1) dynamic infected place tracing strategy and 2) static ranked based strategy.

In the infected place tracing, we traced the infected states, then plan the mitigation strategies according to them. For this type of mitigation strategies, we considered three cases; 1) *case-1: only infected*: applied control only in the infected states; 2) *case-2: infected & first neighbors*: applied control in the infected states with its first neighboring states (whose distance is less than $500km$), and 3) *case-3: infected & first neighbors & second neighbors*: applied control in the infected states with its first neighboring states, and also with its second neighboring states (whose distance is in $500 - 1000km$). For infected tracing control measure, we kept track of infected places monthly. If SNi sub-network is infected for month t, then control measures were applied for the month $t + 1$ based on these three cases.

In the static ranked based mitigation strategy, we ranked the states by different variables (for example, temperature, size of the avian population etc.). For this strategy, we considered three cases; 1) *temp.*: states ranked by temperature, 2) *pop.*: states ranked by avian population size, and 3) *temp. & pop.*: states ranked by temperature and avian population size both, then we applied management measures in the top 30% of the states.

A.7 Architecture of the framework

The ABC-SMC network model selection framework has five major units. The input unit deals with the four sets of data: 1) geographic locations of the contagious United States, 2) American Robin population data for each state, 3) state-level average monthly temperature, and 4) state-level WNV weekly human incidence data. The sub-network generation unit generates a sub-network for each state based on the American Robin population data. The ABC-SMC method estimates a new set of parameters from the prior distribution and selects a network kernel model. The network generation unit will create connections between sub-networks according to the selected network kernel model (a. exponential kernel network model, b. power-law flyway kernel network model, and c. power-law biased by flyway kernel network model). The selected network kernel model was simulated

with the estimated parameters. We used Generalized Epidemic Modeling Framework (GEMF) for simulation. The output of this block is the weekly WNV human cases, which have been used in the ABC-SMC block to compare the simulated result with the actual observed human incidence data.

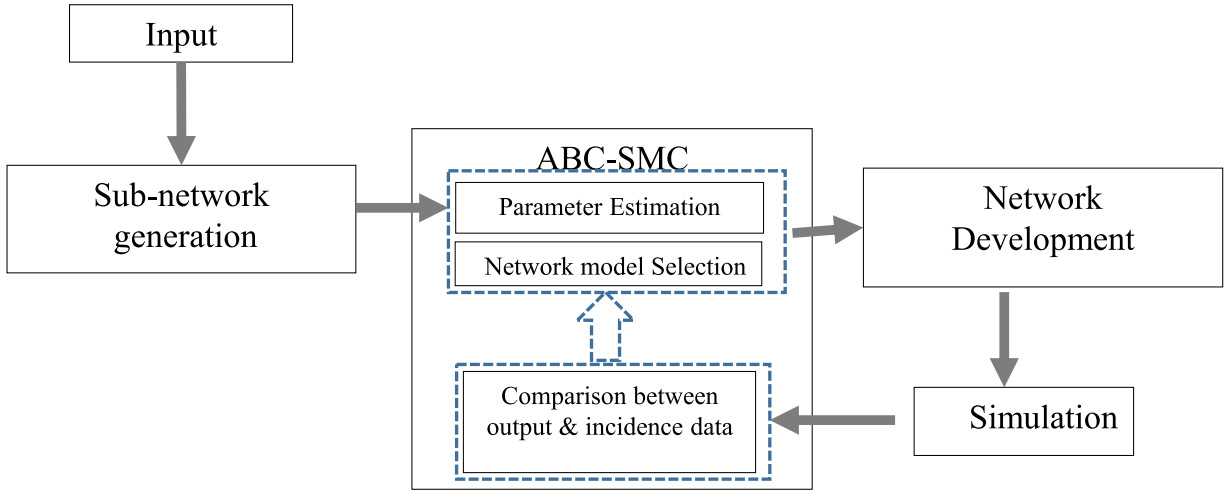


Figure A.6: **Architecture of the ABC-SMC network model selection framework.**

A.8 Results

We developed a novel flexible individual based heterogeneous network framework to test three WNV dispersal kernels across the contiguous United States based on human case data distributions. We used this framework for the year 2014, 2015, and 1016. The results for network formulation, parameter estimation, and dispersal kernels selection using Bayesian inference are given below for the year 2015 and the results for other two years are given in the Moon et al. [11].

A.8.1 Network framework

In this spatial-temporal individual-based heterogeneous network framework, we used three distance kernel models. The fundamental basic WNV epidemic model is the same for all the three network kernels. In the entire network, there are 49 sub-networks representing the 48 adjoining contiguous states plus the District of Columbia. All sub-network nodes are locally connected. The

topology of the local network is *Erdos-Renyi*. The total nodes for the year 2015 was $|V| = 7657$ and the scaling constant is $S_c = 0.02$. Here, $E = E_l \cup E_{dd}$; $|E_l|$ is the number of total intra-links for all local networks, which is around 167000-170000 and $|E_{dd}|$ is the number of total inter-links among sub-networks. We started the epidemic from states with the highest human incidence prior to June. We started the epidemic for the year 2015 by adding two infected nodes, one in sub-network *SN4* (California) and another in sub-network *SN42* (Texas). Connections among sub-networks are developed by distance dispersal kernels. Parameters for these kernels are estimated from the ABC-SMC method.

A.8.2 ABC-SMC for parameter estimation and model comparison

Parameter estimation

ABC-SMC parameter estimation was applied to three dispersal kernel network models separately. For each set of prior distributions, convergence to the posterior distribution was achieved after 13-15 SMC rounds. Convergence of the posterior distributions was monitored by visual inspection of the outputs from consecutive SMC rounds. The prior distribution for exponential network parameter was, $K_e \sim U(0.1, 0.3)$, for power-law $K_{pl} \sim U(2, 4)$, for power-law biased by flyway was $K_{pl} \sim U(2, 4)$. Prior distribution for constant of transmission rate β_0, transition rate from exposed to infectious λ, and human spillover rate η is same for three kernel models; $\beta_0 \sim U(0, 15)$, $\lambda \sim U(0.025, 10)$ and $\eta \sim U(0, 50)$. Perturbation kernels were also uniform, $PK = \alpha U(-1, 1)$, with $\alpha = 0.5(max\theta_{p-1} - min\theta_{p-1})$, here θ_{p-1} is the set of a parameter values in the previous population. We used weekly human case data for 49 locations, as observed data. The estimated parameters for this three dispersal kernel network models for 2015 are presented in Table A.1.

Model comparison

ABC-SMC for model selection allows us to estimate posterior model distributions. We used this algorithm to compare the three distance kernels. Prior distributions and perturbation kernels are the same for both the model selection and the parameter estimation algorithm. Here we used one more prior distribution for discrete model parameter; $m \sim U(1, 3)$. The tolerance vector for ABC-SMC

Table A.1: Estimated parameters for the year 2015 from ABC-SMC parameter estimatio*Estimated using data from the Centers for Disease Control and Prevention (CDC) [3], the National Centers for Environmental Information [4], and Clements et al. [5].

Parameter	Exponential	Power-law	Power-law biased by flyway	Source
Network Parameter, K				
mean	0.1264	3.3844	2.3147	
median	0.1216	3.3924	2.2690	Estimated*
(95% CI)	(0.1235, 0.1294)	(3.3329, 3.4260)	(2.3030, 2.3264)	
Constant for transmission rate, β_0				
mean	0.0439 day^{-1}	0.2026 day^{-1}	0.0059 day^{-1}	
median	0.0362 day^{-1}	0.0526 day^{-1}	0.0061 day^{-1}	Estimated*
(95% CI)	(0.0354, 0.0524 day^{-1})	(0.0574, 0.3478 day^{-1})	(0.0058, 0.0059 day^{-1})	
Transition rate from exposed to infectious node, λ				
mean	0.0884 day^{-1}	0.1069 day^{-1}	0.0721 day^{-1}	
median	0.0823 day^{-1}	0.1059 day^{-1}	0.0706 day^{-1}	Estimated*
(95% CI)	(0.0820, 0.0948 day^{-1})	(0.0940, 0.1197 day^{-1})	(0.0718, 0.0724 day^{-1})	
Bird Recovery rate, δ				
range	0.2-0.25 day^{-1}	0.2-0.25 day^{-1}	0.2-0.25 day^{-1}	[177]
Human spillover, η				
mean	0.2175 day^{-1}	0.2141 day^{-1}	0.4558 day^{-1}	
median	0.2173 day^{-1}	0.2154 day^{-1}	0.4599 day^{-1}	Estimated*
(95% CI)	(0.2098, 0.2252 day^{-1})	(0.2071, 0.2210 day^{-1})	(0.4479, 0.4637 day^{-1})	

model selection algorithm is, $\epsilon = \{2200, 2000, 1800, 1600, 1400, 1200, 1100, 1000\}$. The target and intermediate distributions of model parameters are shown in Fig A.7.

We calculated the Bayes factor from the marginal posterior distribution of m, which we took from the final or last population. In the final population for 2015, exponential distance kernel model ($m = 1$) was selected for 64 times, power-law distance kernel ($m = 2$) was selected for 95 times and power-law influenced by flyway distance kernel model ($m = 3$) was selected for 841 times. Bayes factor $B_{3,1} = 841/64 = 13.1406, B_{3,2} = 841/95 = 8.8526$. In the marginal posterior distribution of three models, there is positive evidence in favor of power-law influenced by flyway distance kernel when compared with other two models [30]. The distribution of parameters for power-law influenced by flyway for 2015 are presented in Fig A.8.

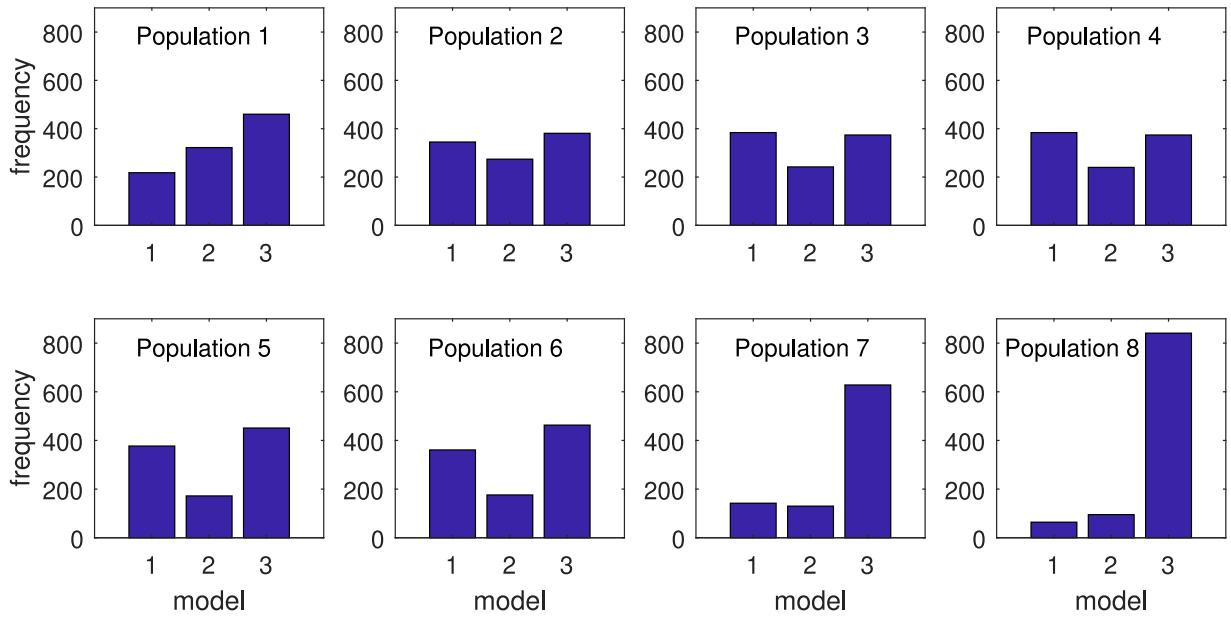


Figure A.7: **Population of the marginal posterior distribution of the three models for the year 2015.** Model-1 represents exponential kernel, model-2 represents power-law kernel, and model-3 represents power-law influenced by flyway kernel. Here, Population-8 is the approximation of the final marginal posterior distribution of model parameter m and population 1-7 are intermediate distributions. Population-0 is the discrete uniform prior distribution, which is not shown here.

A.8.3 Performance of the power-law-flyway network model

To test the performance of this framework, we used estimated parameters from Table A.1 for power-law kernel influenced by flyway. We set the parameters value; $K_{pl} = 2.3147, \beta_0 = 0.0059 day^{-1}$, $\lambda = 0.0721 day^{-1}$, and $\delta = 0.2031 day^{-1}$. The simulation period for the avian population model is from week-23 to week-44. The output of avian population was used as the input of zoonotic spillover compartment. Then we compared the output of zoonotic spillover compartment with human case data for week 24 to week 45. We considered a one-week lag between WNV incidence in birds and WNV incidence in humans. In humans, WNV-infected individuals (approximately 20%) develop a mild febrile illness after 3–6 days [178]. Peak of reporting of dead birds is one week prior than the reporting peak of human incidence [179].

We compared the total yearly incidence of human WNV from this model with the state level reported case data. The detail results are shown in [11]. For 2015, we found that the case data

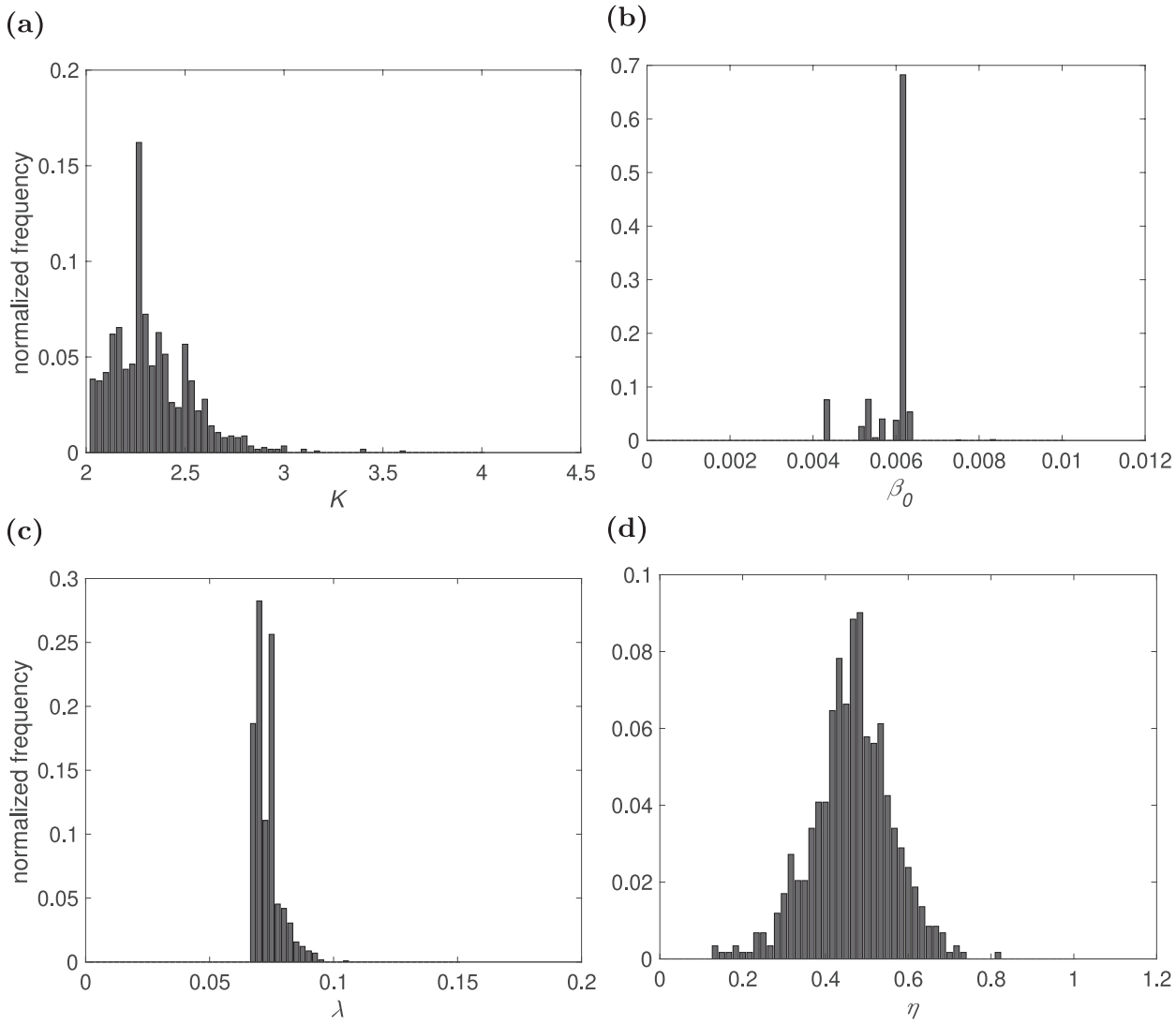


Figure A.8: **Histograms of the approximated posteriors distribution of parameters for power-law influenced by flyway kernel for the year 2015.** a) Network Parameter K ; b) constant for transmission rate β_0; c) transition rate from exposed to infectious node λ, and d) human spillover η.

for 42 of 49 locations were within the simulation results. The states where human cases were different from the simulation results were *over-reported* states (Nevada) and *under-reported* states (Louisiana, Mississippi, Nebraska, North Dakota, South Dakota, and Washington). The possible reason for this mismatch are reporting error or overwintering of virus in birds or mosquitoes or another bird species (not robins) is the key reservoir species for that state

A.9 Summary

We proposed an individual-based heterogeneous network framework and tested three dispersal kernels to understand the spatial spread patterns of WNV human case data across the contiguous United States.

This framework requires fewer parameters and has more flexibility to represent the spatial-temporal dynamics of WNV. Adding parameters can make the framework more realistic, for example, more competent bird species, landscape features for habitat preferences of host and vector species, daylight conditions [155], pathogen invasion from outside of USA, variable susceptibility among different hosts and vectors, WNV strain variability, mosquito and virus overwintering, vertical transmission, human movement characteristics etc.. However, inclusion of too many factors increases model complexity which makes model optimization difficult given the availability of limited observational data. On the other hand, a simple model may insufficient to represent WNV spatial dynamics. Computational models need to be developed and parameters calculated with sufficient detail to be biologically accurate if they are used to evaluate epidemic management measures. However, for most biological systems, reliable parameter information is unknown. Unknown parameters or inaccurate assumptions add uncertainty to the model. Our framework has only four parameters to estimate (network Parameter K, transmission rate β, transition rate from exposed to infectious state, λ, and human spillover, η). This framework has compartments only for the avian population (susceptible, exposed, infected, and recovered), and is not species specific. We reduced the compartments for vector population by implementing them implicitly through transmission rate between infected nodes and susceptible nodes. The presented framework and dispersal kernel network model has an intermediate complexity that approximate Bayesian computation based on sequential Monte Carlo sampling (ABC-SMC) method successfully calibrated and estimated the parameters with the available data. If more data becomes available, it is possible to add them in this model for improved performance of the model.

Furthermore, this framework is flexible and therefore can represent various hosts and vectors including with population seasonality, which plays an important role in WNV dynamics. For host population seasonality, we added a node property *Activity*, this property allows us to control active

host populations in the network in a specific time period. We added vector seasonality in this framework with a temperature dependent transmission rate.

This framework proposed one exponential and two fat-tailed distance kernel models for long-distance transmission of WNV with each model having increasing complexity and similarities to natural avian movement. WNV spatial distribution is very complex because WNV can infect more than 300 bird species, some of which are residential birds and short-distance migrators which disperse less than 500 km distances (short connections) whereas some species are long-distance migratory birds creating long connections. The long-distance migratory birds are the long-distance dispersal (LDD) agents for WNV. Previous studies tried to analyze spreading of WNV using a traveling wave with constant velocity, however, WNV spread more rapidly across the North America than would be expected from the assumption of constant velocity traveling wave[180]. Likely this is because traveling wave models unlike distance dispersal kernel models for WNV spreading do not capture the long-distance migrating birds which can have various migratory ranges and distances. Distance dispersal kernels have more flexibility to represent the different bird migration distances and can account for accelerating invasions. However, exponential kernels produce short-connections and therefore like traveling waves are limited to constant expansion, unlike fat-tailed power-law kernels which can generate accelerating invasions by creating the long-distance connections from migratory birds [181]. However, a general fat-tailed power-law kernel makes long-distance links in every direction which does not follow the incidence of WNV. Instead, a power-law-flyway kernel can be used to produce the long connections in the direction of flyways and short links in other directions. Bayesian inference was used to test which of the three kernel models best described WNV distribution on the network for three most recent years (2014- 2016). The power-law-flyway kernel best described the distribution of WNV cases because the long-range WNV transmission was concentrated mainly along the migratory bird flyways. The general power-law kernel overestimated the incidence data in some states because it was creating long-distance links in all directions.

The performance for the power-law-flyway dispersal kernel model was evaluated for the three most recent years (2014-2016) when WNV was endemic in the USA. The observed case data for the 49 locations were within the range of the simulated results for 41 states for 2014, 42 states for

2015, and 45 states for 2016 [11]. For all three years, the simulated results were similar to the observed data, except in Colorado, Louisiana, Mississippi, Nevada, Nebraska, North Dakota, and Washington. Nevada was *over-reported* for 2015 and all others were *under-reported*. The power law flyway dispersal kernel network model reported more WNV human incidence in Nevada than reported cases, one possible reason for over-reporting cases in Nevada has rural areas, which tend to under report human cases, whereas mosquito control districts and health departments, focused in urban areas, must test birds and mosquitoes, which explains why CDC reported WNV infected mosquitoes in 25% of counties in Nevada. The under-reported states had more human cases than predicted by the model. Under-reporting by the power-law-flyway kernel network model is likely because overwintering of the virus in some states (for example, Louisiana, Mississippi etc.), which was not considered. The overwintering infected *Culex* mosquitoes can stay in hibernacula such as sewers, houses, caves, and other warm areas in urban, suburban, and rural areas and initiate the outbreak in the spring. Furthermore, there may be under-reporting of cases by the model if robins are not the main reservoir species in a state, which would be predicted between gulf coast states (Louisiana and Mississippi) and northern states such as North and South Dakota and Washington.

Stochastic simulations are useful tools to select the optimal future mitigation strategy after outbreaks of invasive species and pathogens. The foot-and-mouth disease (FMD) epidemics in 2001 in the United Kingdom developed by Keeling et al. [59], and mitigation strategies for pandemic influenza in the United States [58] are two well developed models with similarities to the current model. These models explore possible control measures such as culling, vaccination etc. for FMD [59], and vaccination, quarantine etc for influenza [58]. Most of these strategies can be examined with the network framework however, avian culling or vaccination for WNV control is not feasible. Vector control (or mosquito control) is a viable mitigation strategy for WNV, which is not considered by the other two models (FMD and influenza). To be applicable to any pathogens and inclusive of new mitigation methods, the mitigation strategies are non-specific and the predicted effectiveness of the mitigation methods can be adjusted to other methods. In the planning of the mitigation strategies, there is a trade-off between control measures effectiveness and their cost both monetary and loss of life. A stochastic simulation tool can decide the optimal mitigation strategy by dealing with this trade-off.

Mitigation strategies for WNV were tested using the power-law-flyway dispersal kernel network model. The mosquito management measures are not specific to larvae or adults, rather simply generally accepted best practices to reduce mosquito abundance for the purpose of reducing pathogen transmission. The mitigation strategy analysis proposes supplemental measures in addition to the existing mosquito management in each state because the states had yearly reported WNV cases despite the existing management methods. To reduce WNV spread, a theoretical policy would be management in neighboring regions and not exclusively in the infected places. Although this approach can cost more at the beginning of the epidemic season however at the end, it can reduce total cost by decreasing the size of the epidemic. If management measures are applied only in the infected states, it is not possible to control the epidemic because of long-distance migratory birds. This is statewide management in a unified effort. We acknowledge that states do not conduct mosquito management in this way, but to test the spillover it was necessary to do the simulation in this way because only state-level data was available. The infected place tracing mitigation technique has been used to control other diseases (for example, FMD, influenza etc.), although their host population and control measure means are different, however, the main concept behind the mitigation techniques are similar. The findings from this research to control WNV epidemic can be useful to select optimal mitigation strategies for other pathogens.

This research showed that the inclusion of directional long-distance dispersal of migratory birds improves model representations of the spatial patterns of WNV spread in the United States. The simulation of our framework in the context of long-distance directional dispersal suggested that cooperation and communication can facilitate early treatment and reduced outbreak sizes because of reduced WNV dispersal by American robins.

Appendix B

Proofs and examples of chapter 4

B.1 Proof of theorem 1 and derivation of the Q

In this section, we present the derivation of the Q in Eq. (30). To do this, we derive the expression for network-state $G(t+\Delta t)$ when $G(t)$ is given. Here, Δt is a very small time period when only one event can occur. Let, the network-state at any time t be,

$$G(t) = g_Z = g_{z1}(t) \otimes ..... \otimes g_{zC}(t) \tag{B.1}$$

The network-state will change by the one transition in the group-state of group i,

$$E[G(t+\Delta t)|G(t) = g_Z] = \sum_{i=1}^{C} g_{z1}(t) \otimes ..... \otimes E[g_i(t+\Delta t)|G(t) = g_Z] \otimes ... \otimes g_{zC}(t) \tag{B.2}$$

The expression for the conditional expectation of a group $E[g_i(t+\Delta t)|G(t) = g_Z]$ can get from the Eq. (25) as,

$$E[g_i(t+\Delta t)|G(t) = g_Z] = \sum_{q=1}^{q_n} \Delta_{i,\delta_q}^T g_{zi}(t)\Delta t + \sum_{q=1}^{q_e} (\sum_{j=1}^{C} \mathcal{A}_g(i,j)x_{j,r})\Delta_{i,\beta_q}^T g_{zi}(t)\Delta t + g_{zi}(t) + o(\Delta t) \tag{B.3}$$

Now, from the definition of expectation and the law of total probability, we get the network-state at $t + \Delta t$ time,

$$E(G(t + \Delta t)) = \sum_{Z} E[G(t + \Delta t)|G(t) = g_Z]Pr[G(t) = g_Z] \quad \text{(B.4)}$$

Here, the range of Z for the summation in Eq. (B.4) is $1 : \binom{N_1+M-1}{M-1}\binom{N_2+M-1}{M-1}....\binom{N_C+M-1}{M-1}$.
From Eq. (B.3) and (B.4),

$$E(G(t + \Delta t)) = \sum_{q=1}^{q_n} Q_{\delta_q}^T E[G(t)]\Delta t + \sum_{q=1}^{q_e} Q_{\beta_q}^T E[G(t)]\Delta t + E[G(t)] + o(\Delta t) \quad \text{(B.5)}$$

Here,

$$Q_{\delta_q} = \sum_{i=1}^{C} I_{\binom{N_1+M-1}{M-1}\times\binom{N_1+M-1}{M-1}} \otimes .... \otimes \Delta_{i,\delta_q} \otimes ..... \otimes I_{\binom{N_C+M-1}{M-1}\times\binom{N_C+M-1}{M-1}} \quad \text{(B.6)}$$

The Z^{th} column of Q_{β_q} is,

$$Q_{\beta_q}(:, Z) = \sum_{i=1}^{C} g_{z1}(t) \otimes ..... \otimes \left(\sum_{j=1}^{C} \mathcal{A}_g(i, j)x_{j,r}\right)\Delta_{i,\beta_q} g_{zi}(t) \otimes .... \otimes g_{zC}(t) \quad \text{(B.7)}$$

Let,

$$Q = \sum_{q=1}^{q_n} Q_{\delta_q} + \sum_{q=1}^{q_e} Q_{\beta_q} \quad \text{(B.8)}$$

We will get differential equations for the underlying continuous-time Markov process for the group-based approach from Eq. (B.5) by letting $\Delta t \to 0$,

$$\frac{d}{dt}E[G] = Q^T E[G] \quad \text{(B.9)}$$

B.2 Proof of theorem 3

The first-order moment closure approximation inside a group allows us to consider that the nodes in any two compartments in a group i are uncorrelated. Therefore, $Cov[x_{i,m}x_{i,n}] \approx 0$. The epidemic spreading process is not Markovian anymore. A set of M differential equations can describe the time evolution of the expected value of the population of a group for an M compartmental epidemic model. If a node of the group i move from compartment m to n with rate δ_q then the population of the compartment m and n will change with rate δ_q. An $M \times M$ transition rate matrix contains the transition rate between two compartments. The Eq. 4.28 has similarity with the individual-based GEMF mean-field equation except the group-based part. Please check details in [1].

B.3 Intra- and inter-group mean-field equations for the SIS, SIR, and SEIR epidemic models

B.3.1 Susceptible-infected-susceptible (SIS)

The SIS model [14] has two types of transitions; one an edge transition (susceptible-to-infected) and the other a nodal transition (infected-to-susceptible). The infected compartment is the influencer compartment for the edge transition. The mean-field equation for the group-based framework of the SIS epidemic model can be written as

$$\begin{bmatrix} \dot{\rho_{i,S}} \\ \dot{\rho_{i,I}} \end{bmatrix} = \Big(\sum_{j=1}^{C} \frac{L_{ij}}{\mathcal{N}_i} \rho_{j,I} \Big) \underbrace{\begin{bmatrix} -\beta & \beta \\ 0 & 0 \end{bmatrix}^T}_{\Phi_\beta^T \text{ matrix}} \begin{bmatrix} \rho_{i,S} \\ \rho_{i,I} \end{bmatrix} + \underbrace{\begin{bmatrix} 0 & 0 \\ \delta & -\delta \end{bmatrix}^T}_{\Phi_\delta^T \text{ matrix}} \begin{bmatrix} \rho_{i,S} \\ \rho_{i,I} \end{bmatrix} \quad \text{(B.10)}$$

Here, $\rho_{i,S}$ and $\rho_{i,I}$ represent the fraction of susceptible and infected nodes in the group i and $\rho_{i,S} + \rho_{i,I} = 1$ at any time t. The first and second parts in the Eq. (B.10) are for the edge transition $S \to I$ (susceptible-to-infected) and nodal transition $I \to S$ (infected-to-susceptible), respectively. The rate for edge transition is β and the rate for nodal transition is δ. This process has $(2-1)C$

ordinary differential equations.

B.3.2 Susceptible-infected-recovered (SIR)

The SIR epidemic spreading has three compartments and two types of transitions, one an edge transition (susceptible-to-infected) and the other a nodal transition (infected-to-recovered). The infected compartment is the influencer compartment of the edge transition. The mean-field approximation of the susceptible-infected-recovered (SIR) epidemic model for the individual-based framework is developed by Sharkey et al. [182]. Here, we present the equation for the group-based framework as

$$\begin{bmatrix} \dot{\rho_{i,S}} \\ \dot{\rho_{i,I}} \\ \dot{\rho_{i,R}} \end{bmatrix} = \Big(\sum_{j=1}^{C} \frac{L_{ij}}{\mathcal{N}_i} \rho_{j,I} \Big) \underbrace{\begin{bmatrix} -\beta & \beta & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}^T}_{\Phi_\beta^T \text{ matrix}} \begin{bmatrix} \rho_{i,S} \\ \rho_{i,I} \\ \rho_{i,R} \end{bmatrix} + \underbrace{\begin{bmatrix} 0 & 0 & 0 \\ 0 & -\delta & \delta \\ 0 & 0 & 0 \end{bmatrix}^T}_{\Phi_\delta^T \text{ matrix}} \begin{bmatrix} \rho_{i,S} \\ \rho_{i,I} \\ \rho_{i,R} \end{bmatrix} \quad \text{(B.11)}$$

The first part of this equation is for the transition $S \rightarrow I$ and the second part is for the transition $I \rightarrow R$. The number of non linear differential equations for the mean-field approximation of the SIR epidemic model for the group-based framework is $(3-1)C$ as at any time t, $\rho_{i,S} + \rho_{i,I} + \rho_{i,R} = 1$.

B.3.3 Susceptible-exposed-infected-recovered (SEIR)

The SEIR epidemic model has four compartments and three transitions: susceptible-to-exposed, exposed-to-infected, and infected-to-recovered. The first transition is the edge transition, where the transition rate β is influenced by the number of infected nodes in the neighboring groups. The other two transitions are nodal transitions with the rate δ_1 ($E \rightarrow I$) and δ_2 ($I \rightarrow R$), respectively.

The group-based mean-field equation for the SEIR epidemic is given below:

$$\begin{bmatrix} \dot{\rho}_{i,S} \\ \dot{\rho}_{i,E} \\ \dot{\rho}_{i,I} \\ \dot{\rho}_{i,R} \end{bmatrix} = \Big(\sum_{j=1}^{C} \frac{L_{ij}}{N_i} \rho_{j,I} \Big) \underbrace{\begin{bmatrix} -\beta & \beta & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}^T}_{\Phi_\beta^T \text{ matrix}} \begin{bmatrix} \rho_{i,S} \\ \rho_{i,E} \\ \rho_{i,I} \\ \rho_{i,R} \end{bmatrix} + \underbrace{\begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & -\delta_1 & \delta_1 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}^T}_{\Phi_{\delta_1}^T \text{ matrix}} \begin{bmatrix} \rho_{i,S} \\ \rho_{i,E} \\ \rho_{i,I} \\ \rho_{i,R} \end{bmatrix} + \underbrace{\begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & -\delta_2 & \delta_2 \\ 0 & 0 & 0 & 0 \end{bmatrix}^T}_{\Phi_{\delta_2}^T \text{ matrix}} \begin{bmatrix} \rho_{i,S} \\ \rho_{i,E} \\ \rho_{i,I} \\ \rho_{i,R} \end{bmatrix} \tag{B.12}$$

The first part of the Eq. (B.12) represents the edge transition $S \rightarrow E$, the second part represents the nodal transition $E \rightarrow I$, and the last part represents the nodal transition $I \rightarrow R$. The number of nonlinear ordinary differential equations for this epidemic model is $(4-1)C$, as at any time t, $\rho_{i,S} + \rho_{i,E} + \rho_{i,I} + \rho_{i,R} = 1$.

B.4 Simulation results in an Erdös-Rényi (ER) random network

In this section, we compare the simulation results from the group-based approaches with the individual-based approaches in the Erdös-Rényi (ER) random network. Stochastic numerical simulations of the exact continuous-time Markov process of the individual-based approach is the benchmark of this comparison. In this section, we use a Erdös-Rényi (ER) network with ($N = 10000, p = 0.01$). The simulation results for SIS ($\beta = 0.0167, \delta = 1$), and SIR ($\beta = 0.0167, \delta = 1$) epidemic spreading are presented in the Fig 4.4, and4.6, respectively. In Fig 4.4 and 4.6, sub-plot (a) is presenting simulation results from numerical simulations of the continuous-time Markov

process of the individual-based approach, sub-plot (b) is presenting simulation results from the mean-field individual-based NIMFA model, sub-plots (c)-(e) is presenting simulation results from the mean-field approximation of the group-based GgroupEM model, and sub-plot (f) is presenting the merging of sub-plots (a)-(e). A comparison of simulation time between individual-based and group-based approaches of Fig 4.4 and 4.6 is given in Table B.1.

Table B.1: Comparison of simulation time between individual-based and group-based approaches.

case	No. of Groups	simulation time	
		SIS	SIR
Individual-based stochastic	-	3060.6s	380.038s
Individual-based mean-field	-	35.352s	17.582s
group-based mean-field	100	0.334s	0.149s
group-based mean-field	10	0.0123s	0.0168s

From Fig 4.4 and 4.6, the group-based approach can produce the similar dynamics as the individual-based approach in SIS and SIR disease spreading in a Erdös-Rényi (ER) random network. From Table B.1, the simulation time for group-based approaches is less than the simulation time for the individual-based approaches.

B.5 An example of the group-based epidemic model

A network with $N = 5$ nodes. The nodes are divided into $C = 2$ groups. The first group has $N_1 = 2$ nodes and the second group has $N_2 = 3$. For a susceptible-infected-susceptible (SIS) epidemic process, the first group has $\binom{N_1+M-1}{M-1} = 3$ states and the second group has $\binom{N_2+M-1}{M-1} = 4$ states. The

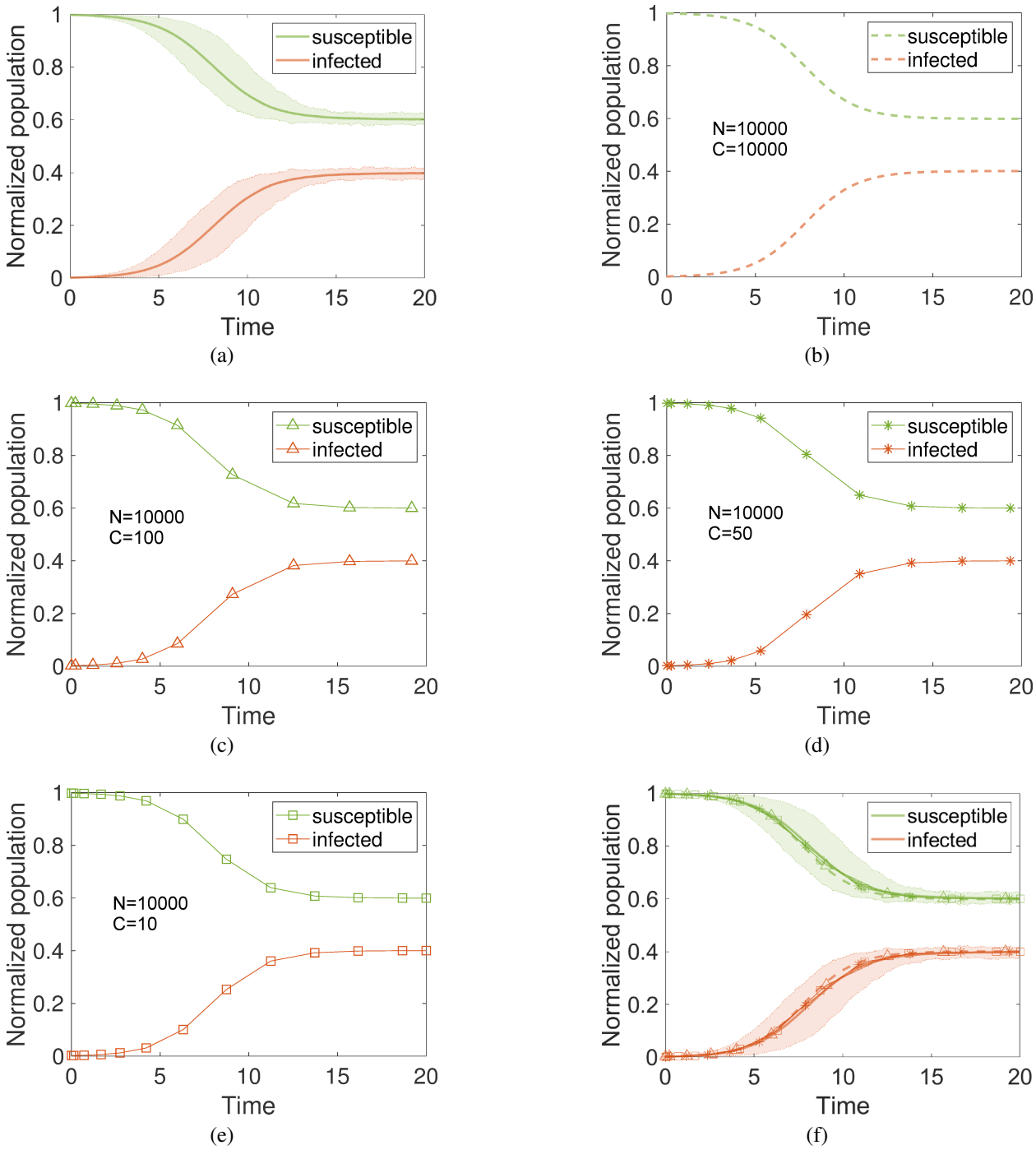


Figure B.1: **Time dynamics for an SIS epidemic in the Erdös-Rényi (ER) random network** ($N = 10000, p = 0.01$)**;** a) Stochastic numerical simulation of the exact continuous-time Markov process of the individual-based approach; solid lines represent the average of the 1000 simulations and shaded areas represent the region of the stochastic simulations; b) Individual-based: $N = C = 10000, \mathcal{N}_1 = \mathcal{N}_2 = ... = \mathcal{N}_C = 1$, simulation time = $35.352s$; c) group-based: $C = 100, \mathcal{N}_1 = \mathcal{N}_2 = ... = \mathcal{N}_C = 100$, simulation time = $0.334s$; d) group-based: $C = 10, \mathcal{N}_1 = \mathcal{N}_2 = ... = \mathcal{N}_C = 200$, simulation time = $0.153s$; e) group-based: $C = 10, \mathcal{N}_1 = \mathcal{N}_2 = ... = \mathcal{N}_C = 1000$, simulation time = $0.0123s$; and f) merging of all sub-plots a-e.

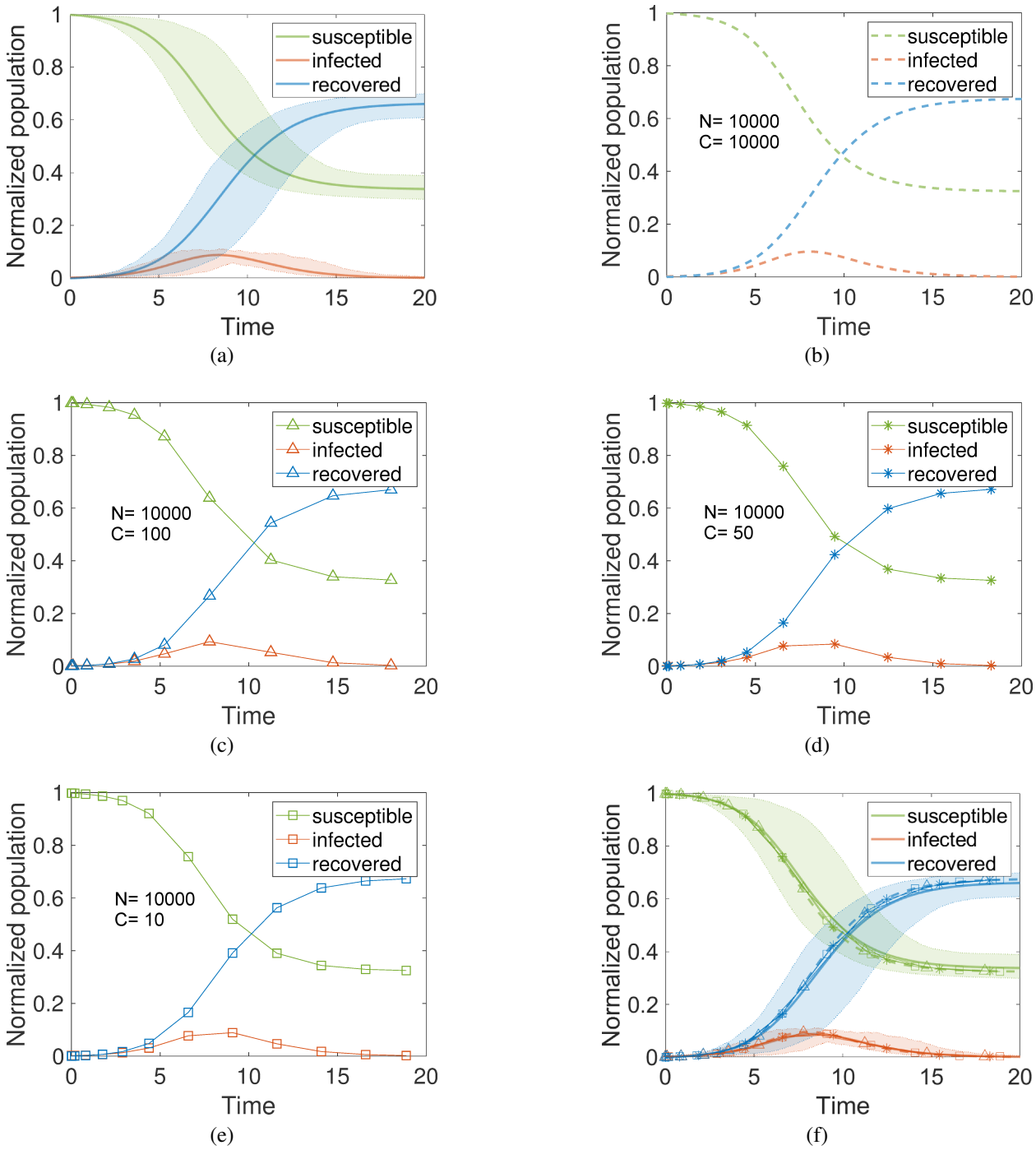


Figure B.2: **Time dynamics for an SIR epidemic in the Erdös-Rényi (ER) random network** ($N = 10000, p = 0.01$)**;** a) Stochastic numerical simulation of the exact continuous-time Markov process of the individual-based approach; solid lines represent the average of the 200 simulations and shaded areas represent the region of the stochastic simulations; b) Individual-based: $N = C = 10000, \mathcal{N}_1 = \mathcal{N}_2 = ... = \mathcal{N}_C = 1$, simulation time = $17.582s$; c) group-based: $C = 100, \mathcal{N}_1 = \mathcal{N}_2 = ... = \mathcal{N}_C = 100$, simulation time = $0.149s$; d) group-based: $C = 50, \mathcal{N}_1 = \mathcal{N}_2 = ... = \mathcal{N}_C = 200$, simulation time = $0.091s$; e) group-based: $C = 10, \mathcal{N}_1 = \mathcal{N}_2 = ... = \mathcal{N}_C = 1000$, simulation time = $0.0168s$;and f) merging of all sub-plots a-e.

description of the group-states are given below

$$group_1 = \begin{bmatrix} - & - & - & V_1 \\ | & * & * & [0,2] \\ * & | & * & [1,1] \\ * & * & 1 & [2,0] \end{bmatrix}$$

$$group_2 = \begin{bmatrix} - & - & - & - & V_2 \\ | & * & * & * & [0,3] \\ * & | & * & * & [1,2] \\ * & * & | & * & [2,1] \\ * & * & * & | & [3,0] \end{bmatrix} \quad \text{(B.13)}$$

The SIS epidemic process has two compartments: susceptible and infected. One divider '|' can divide nodes '$*$' into two compartments. At first, we will present the steps to get Q_{δ_1} for the nodal transition infected-to-susceptible, then we will present the steps to get Q_{β_1} for the edge transition susceptible-to-infected.

Here, the *transition-specific matrix* for group-1 and group-2 for the nodal transition from susceptible-to-infected compartment will be

$$\Delta_{1,\delta_1} = \begin{bmatrix} -2\delta & 2\delta & 0 \\ 0 & -\delta & \delta \\ 0 & 0 & 0 \end{bmatrix}$$

$$\Delta_{2,\delta_1} = \begin{bmatrix} -3\delta & 3\delta & 0 & 0 \\ 0 & -2\delta & 2\delta & 0 \\ 0 & 0 & -\delta & \delta \\ 0 & 0 & 0 & 0 \end{bmatrix} \quad \text{(B.14)}$$

The definition of the *transition-specific matrix* is given in Eq. 4.13.

$$Q_{\delta_1} = \begin{bmatrix} -5\delta & 3\delta & 0 & 0 & 2\delta & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & -4\delta & 2\delta & 0 & 0 & 2\delta & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & -3\delta & \delta & 0 & 0 & 2\delta & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -2\delta & 0 & 0 & 0 & 2\delta & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & -4\delta & 3\delta & 0 & 0 & \delta & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & -3\delta & 2\delta & 0 & 0 & \delta & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & -2\delta & \delta & 0 & 0 & \delta & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & -\delta & 0 & 0 & 0 & \delta \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -3\delta & 3\delta & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -2\delta & 2\delta & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -\delta & \delta \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \tag{B.15}$$

The Q_{δ_1} matrix can be obtained for this case from Eq. (B.6), which is presented in Eq (B.15).

Let the states of the groups at time t, be $g_{z1}(t) = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}$, and $g_{z2}(t) = \begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \end{bmatrix}$. Therefore, the group-state at time t is $G(t) = \begin{bmatrix} 0,0,0,0,0,0,1,0,0,0,0,0 \end{bmatrix}^T$.

Now, the *transition-specific matrix* for the edge transition susceptible-to-infected with the rate β

is

$$\Delta_{1,\beta_1} = \begin{bmatrix} 0 & 0 & 0 \\ \beta & -\beta & 0 \\ 0 & 2\beta & -2\beta \end{bmatrix}$$

$$\Delta_{2,\beta_1} = \begin{bmatrix} 0 & 0 & 0 & 0 \\ \beta & -\beta & 0 & 0 \\ 0 & 2\beta & -2\beta & 0 \\ 0 & 0 & 3\beta & -3\beta \end{bmatrix} \quad \text{(B.16)}$$

Now, the Q_{β_1} matrix for this case can be obtained from Eq. (B.7),

$$Q_{\beta_1}(:, Z) = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ -2\beta\left(\sum_{j=1}^{2} \mathcal{A}_g(2,j)x_{j,2}\right) - \beta\left(\sum_{j=1}^{2} \mathcal{A}_g(1,j)x_{j,2}\right) \\ 3\beta\left(\sum_{j=1}^{2} \mathcal{A}_g(2,j)x_{j,2}\right) \\ 0 \\ 0 \\ 2\beta\left(\sum_{j=1}^{2} \mathcal{A}_g(1,j)x_{j,2}\right) \\ 0 \end{bmatrix} \quad \text{(B.17)}$$