2024

Probabilistic modeling of disease: Addressing uncertainties in within-host and population-level dynamics

Mariah Boudreau
University of Vermont

Probabilistic modeling of disease: Addressing uncertainties in within-host and population-level dynamics

A Dissertation Presented

by

Mariah Cecile Boudreau

to

The Faculty of the Graduate College

of

The University of Vermont

In Partial Fulfillment of the Requirements
for the Degree of Doctor of Philosophy
Specializing in Mathematical Science

October, 2024

Defense Date: July 24th, 2024
Dissertation Examination Committee:

Christopher M. Danforth, Ph.D., Advisor
Laurent Hébert-Dufresne, Ph.D., Advisor
Sarah A. Nowak, Ph.D., Chairperson
Jean-Gabriel Young, Ph.D.
Peter Sheridan Dodds, Ph.D.
Holger Hoock, DPhil, Dean of the Graduate College

Abstract

Mathematical modeling of disease dynamics provides powerful tools to understand, predict, and evaluate emerging diseases. These insights aid public health officials, along with other modelers. With a plethora of models to choose from, it is important to consider a model that encapsulates the stochastic nature of disease dynamics. Stochasticity not only conveys chances of stochastic extinction, but provides probabilistic outcomes, essential for capturing the stochastic nature of the real world. In this thesis, three stochastic models are presented, each addressing uncertainties in mechanisms and interpretation of these models, to aid other modelers and decision makers.

Starting with the source of infection spread, we address the uncertainties in human papillomavirus (HPV) within-host dynamics. The motivation behind this mechanistic cellular dynamics model is that HPV infection progression information, viral load and extinction are not well documented for model inputs. Through a master equation approach, we establish extinction probability, persistence, and viral load metrics from moment information to inform population-level model parameters. Furthermore, the structure of the skin layer, possibly indicating an older individual, impacts differing viral load information and disease propagation.

The subsequent topic of this thesis models interventions on stochastic branching processes on disease spread contact patterns. From an extension of a temporal probability generating function approach, we evaluate different interventions, resulting in cumulative count probability distributions. When comparing intervention strategies, probability distribution output makes comparison difficult. Nevertheless, we establish several metrics to compare these temporal and probabilistic forecasts, providing clear definition on what a decision maker may want to mitigate.

Lastly, the final chapter addresses uncertainty in the giant component analysis application of probability generating functions (PGFs): a sensitivity analysis of the polynomial roots. The condition number of these roots can be evaluated when small perturbations are applied to the coefficients. Two probability distributions are presented as case studies to assess which systems may be more prone to giant component variation, or in the context of disease modeling, final outbreak size variation. This not only evaluates the sensitivity of PGF applications for the first time, but establishes a way to examine the sensitivity of a branching process for other applications.

This thesis explores the uncertainties for disease progression outcomes, case count distribution comparison, and branching process final outbreak size sensitivity. Each chapter contributes method expansion or evaluation, along with commentary on declaring clear assumptions for each method.

Citations

Material from this dissertation has been published in the following form:

Boudreau, M. C., Allen A. J., Roberts, N. J., Allard, A., & Hébert-Dufresne, L.. (2023). *Temporal and probabilistic comparisons of epidemic interventions.* Bull. of Math. Biol., 85(12), 118.

Allen, A. J., Boudreau, M. C., Roberts, N. J. Allard, A., & Hébert-Dufresne, L.. (2022). *Predicting the diversity of early epidemic spread on networks.* Phys. R. Research, 4(1), 013123.

Stuart, R. M., Cohen, J. A., Kerr, C. C., Abeysuriya, R. G., Zimmermann, M., Rao, D. W., Boudreau, M. C., Serin, L., Luojun Y.,& Klein, D. J.. (2024) *Hpvsim: An agent-based model of hpv transmission and cervical disease.* PLOS Comp. Biol., 20(7), e1012181.

Other materials from this dissertation have been submitted for publication:

Boudreau, M. C., Cohen, J. A., & Hébert-Dufresne, L.. (2024). *Within-host infection dynamics with master equations and the method of moments: A case study of human papillomavirus in the epithelium.* Manuscript currently under preparation.

Boudreau, M. C., Thompson, W., Danforth, C. M., Young, J.-G., & Hébert-Dufresne, L.. (2024).*Sensitivity analysis of stochastic polynomial roots, and its application to epidemic forecasting and random graphs.* Manuscript currently under preparation.

Human beings have a remarkable ability to accept the abnormal and make it normal.

-Andy Weir, *Project Hail Mary*

Learning anything takes time, practice, and vulnerability.

-Tori Dunlap, *Financial Feminist*

Acknowledgements

I would not be in this position if it were not for the amazing communities I am a part of. So, from the bottom of my heart, I appreciate the support from those in and out of my academic sphere before and during my graduate studies.

First of, I would like to thank my advisors Laurent Hébert-Dufresne and Chris Danforth for their support, encouragement, and enthusiasm in helping me become a better scientist and human. Laurent's excitement for science is unmatched and it is inspiring to be a student of his. Chris's wisdom through this program has made the difficulties not as heavy and joys much more vibrant.

To my co-authors, Andrea J. Allen, and Nicholas J. Roberts, thank you for banding together on two peer-review publications. To Jamie A. Cohen and the team at the Bill and Melinda Gates Foundation, thank you for letting me join the inspiring workspace of IDM for a summer. To Will Thompson, thank you for assisting with the final push on a project that has been my white whale since the first semester of my graduate studies.

To Josh Minot, John Meluso, Milo Trujilo, Nicholas Landry, Bryn Loftness, Brad Demarest, Juniper Lovato, Sarah Tabor, Melissa Rubinchuk and the QuEST community, your friendships and camaraderie in the office and on campus always made me excited to come to work. To Jean-Gabriel Young and Antoine Allard, thank you for providing support on research and I hope to continue collaborations in the future.

To my parents, Seth and Mallory, your unwavering support means the world, even if there were many jokes about my nerdiness over the years. To Ryan, thank you for telling me I am doing a great job, no matter the level of confusion on my face when looking at my computer.

Table of Contents

Epigraph . . . iii
Acknowledgements . . . iv
List of Figures . . . ix
List of Tables . . . x

Foreword **1**

1 Differential Equation Models of Disease Dynamics **3**
1.1 Mathematical Models of Disease Dynamics . . . 4
1.2 Stochastic Differential Equation Models of Disease Dynamics . . . 5
1.2.1 Stochastic Processes . . . 6
1.2.2 Stochastic Simulation . . . 8
1.2.3 Master Equations . . . 9

2 Within-host infection dynamics with master equations and the method of moments: A case study of human papillomavirus in the epithelium **11**
Abstract . . . 12
2.1 Introduction . . . 12
2.2 Methods . . . 17
2.2.1 Assumptions . . . 17
2.2.2 Master equations . . . 18
2.2.3 Method of moments (MoM) . . . 20
2.2.4 Extinction and non-extinction events . . . 22
2.2.5 Viral load . . . 23
2.2.6 Stochastic simulation algorithm . . . 24
2.3 Application to structured epithelium dynamics . . . 25
2.3.1 Three-cell-type system . . . 25
2.3.2 Five-cell-type system . . . 28
2.4 Results . . . 30
2.4.1 Extinction probability . . . 33
2.4.2 Persistent infections . . . 34
2.4.3 Cumulative virions . . . 35
2.5 Discussion . . . 37

3 Branching Process Models of Disease Dynamics **41**
3.1 Branching processes of disease spread . . . 41
3.2 Percolation on Contact Networks . . . 42

3.2.1 Network Definitions . . . 42
3.2.2 Percolation on Contact Networks . . . 43
3.3 Probability generating functions . . . 46
3.3.1 Generational Spread Analysis . . . 52
3.3.2 Giant Component Analysis . . . 57

4 Temporal and probabilistic comparisons of epidemic interventions 62
Abstract . . . 63
4.1 Introduction . . . 63
4.2 Theoretical Analysis . . . 66
4.2.1 Assumptions . . . 66
4.2.2 Noël et al. probability generating function (PGFs) formalism . 69
4.2.3 Formalism extension: altering transmission . . . 73
4.3 Interventions . . . 76
4.3.1 Uniform or random interventions . . . 78
4.3.2 Targeted network interventions . . . 80
4.3.3 Validation via simulations . . . 83
4.3.4 Comparison of interventions . . . 84
4.4 Case study: Random vs targeted vaccination . . . 86
4.5 Discussion . . . 89

5 Sensitivity analysis of stochastic polynomial roots, and its application to epidemic forecasting and random graphs 92
Abstract . . . 93
5.1 Introduction . . . 93
5.2 Methods . . . 95
5.2.1 Assumptions . . . 95
5.2.2 Probability generating functions . . . 95
5.2.3 Statistical condition of polynomial roots . . . 98
5.3 Case Studies . . . 100
5.3.1 Negative binomial simulations . . . 101
5.3.2 Erdös-Rényi graph simulations . . . 101
5.4 Results . . . 102
5.4.1 Negative binomial: small outbreak sensitivity . . . 102
5.4.2 Erdös-Rényi graphs: sensitive thresholds . . . 104
5.5 Discussion . . . 106

6 Conclusion 109
6.1 Methodological assumptions . . . 110
6.2 Parameter literature review . . . 111

6.3 Future multi-scale model 112
6.4 Final thoughts 114

Bibliography **114**

Appendix **123**

List of Figures

1.1 SIR model dynamics 6
1.2 Probability flow for master equations: 10

2.1 Cervical epithelium 16
2.2 Three-cell-type model schematic 26
2.3 Five-cell-type model schematic 28
2.4 Average and variance evolution of the three-cell-type system 32
2.5 Average and variance evolution of the five-cell-type system 33
2.6 Cumulative probability of extinction of the basal cells for the three and five-cell-type system 35
2.7 Average basal cells of persistent infections for the three and five-cell-type system 36
2.8 Average cumulative virions shed for the three and five-cell-type system 37

3.1 Basic random network 44
3.2 Contact degree distribution, $G_0(x)$, visual representation 46
3.3 Excess degree distribution, $G_1(x)$, visual representation 49
3.4 Second neighbor degree distribution visual representation 50
3.5 Transmission on $G_0(x)$ visual representation 51
3.6 Generational infections 53
3.7 Time evolution of epidemics on a power-law network 56
3.8 Finite component size distribution from a randomly chosen edge, $H_1(x)$ visual representation 57
3.9 Finite component size distribution from a randomly chosen vertex, $H_0(x)$ visual representation 58
3.10 Solving $G_1(u) - u$ for polynomial roots 61

4.1 Schematic of generations of infection through a network with interventions. 66
4.2 Mapping continuous-time dynamics to branching process generations. 67
4.3 Random and targeted rollout comparison and validation. 77
4.4 Flat distributions at generation 10. 80
4.5 Varying targeted vaccination metrics compared to two random vaccination metrics. 87

5.1 PGF root value and outbreak sizes: 100
5.2 Negative binomial condition numbers 104

5.3 Negative binomial condition numbers overlaid with polynomial root and outbreak size 105
5.4 Percolated Erdös-Rényi 106

6.1 Visual Contributions 114

List of Tables

2.1 Three-cell-type system parameters 27
2.2 Five-cell-type system parameters 30

5.1 Table of various diseases 107

Foreword

Upon each inquiry of the details of my graduate studies, I am reminded that mathematics is a language the transcends all subject areas. This universality allows mathematical modeling tools to be extended, adjusted, and applied to various phenomena. With my interest landing on disease dynamics, understanding the nature of these dynamics through a stochastic modeling lens resulted in this thesis. The stochastic models presented in this thesis not only extend particular methodologies, but recognizes uncertainties that other modelers, and decision makers encounter. Due to their probabilistic nature, stochastic models provide an excellent tool to explore and consider these uncertainties.

Each chapter explains an uncertainty, and details a model that aims to fill a gap in current knowledge. Analogous to how disease spread begins, Chapter 2 focuses on the disease dynamics for a within-host application. For a particular infection, we identify knowledge gaps in the infection progression outcomes. Therefore, we establish a framework for cellular infection dynamics. The model outputs, extinction, persistence, and viral load distributions, vary depending on the structure of the skin. Thus this framework accounts for heterogeneity in within-host model outputs, which can be translated into population-level model inputs.

Moving from within-host modeling to population-level modeling, Chapter 4 extends a temporal probability generating function analysis to evaluate intervention strategies. When analyzing two different types of intervention strategies, probability distributions are difficult to compare. This chapter proposes a comparison method for the output distributions. Therefore, decision makers can decide what the goal of the intervention is, then focus on that particular metric from the output.

Finally, Chapter 5 evaluates the sensitivity of final outbreak sizes using probability generating functions. Rather than modeling projections of disease spread, this framework derives the variation in outbreak sizes for perturbed systems. We discuss the implications for other probability generating function sensitivity analysis applications at the end of this chapter. In order to give accurate modeling outcomes, the bounds and scope of a model must be tested. While this has the potential to point out the flaws of a model, all models have flaws and the more we understand how those effect their outcomes, the better.

The concept presented in Chapter 2 follows from work with the Institute for Disease Modeling at the Bill and Melinda Gates Foundation [91]. Chapters 2 and 5 of this thesis are manuscripts under preparation [16, 17]. While Chapter 4 is a peer-reviewed publication, it follows from another publication within our group [4], with more details found in Sec. 3.3.1.

Chapter 1

Differential Equation Models of Disease Dynamics

Diseases have been around long before mathematics. We consistently use mathematical models to understand how the world around us affects or will affect individual and population-level well-being. With the spectrum of mathematical models, we hold the ability to predict, mitigate, explore, and comprehend the complex systems that we interact with each day. In the context of disease dynamics, predicting and mitigating disease progressions can save lives, while exploring and comprehending diseases provides insights into the drivers and macroscopic effects on their dynamics. Given these statements, no wonder a myriad of models came to light during the COVID-19 pandemic.

The tool shed of mathematical disease modeling[1] has grown and diversified over time from influence of various disciplines. The disciplines that have influenced the models of this thesis include biology, statistical physics and combinatorics. When

[1]The box is not big enough anymore.

considering models to address a disease dynamics problem, certain questions will filter out some models. These questions can range from: Is the desired outcome the average of the system? What happens temporally? Or, what about the role of randomness? Some models that address these questions range from mean-field models that provide the average state of the system over time, to agent-based models that provide granularity and heterogeneity through tracking all individual agents. Since one mathematical model cannot address all possible questions, assumptions and desired outputs, it is up to the modeler to clearly define those aspects of their model choice. Nevertheless, this thesis will focus on models that contain a probabilistic aspect. Now, let's consider some of the original models that established the tool shed.

1.1 Mathematical Models of Disease Dynamics

The first recorded instance of mathematical disease modeling was conducted by Daniel Bernoulli in the 1700s. Bernoulli modeled the spread of smallpox in a population using probabilities of death and transmission, along with the effect of immunized people in a population [13]. This model took steps towards defining the variation in susceptible individuals over time, represented as a differential equation [9]. Kermack and McKendrick catapulted the differential equation model further between 1927 and 1933 [48, 50, 49]. These authors most famously established the SIR model, which paved the way for the generalized compartmental model. In the SIR model, each compartment defines a category of individuals, susceptible, $S(t)$, infected, $I(t)$,

and recovered or removed, $R(t)$. The transitions between each compartment allow for temporal changes to occur in each compartment's population. Compartmental models derive the average state of the system each time they are computed. This establishes them as a mean-field model. These rates of change define parameters that include the infection rate β, and the recovery or removal rate, γ. The governing equations for these dynamics are given by,

$$\begin{aligned} \frac{d}{dt}S(t) &= -\beta\frac{SI}{N}, \\ \frac{d}{dt}I(t) &= \beta\frac{SI}{N}, -\gamma I, \\ \frac{d}{dt}R(t) &= \gamma I. \end{aligned} \tag{1.1}$$

Figure 1.1 illustrates the temporal change in each compartment population. While this is the original version of the compartment model, variations have emerged, such as the SIS (susceptible, infected, susceptible) [45]. The flexibility of these compartments and the differential equation structure make defining the average state of a system with categories of individuals easy.

1.2 Stochastic Differential Equation Models of Disease Dynamics

What happens if an infected individual/cell/particle recovers before infecting others? Deterministic mean-field models do not consider this question, and produce their determined dynamics for a specific set of parameters without a stochastic aspect. However, there is potential for a spreading process to never spread past its early

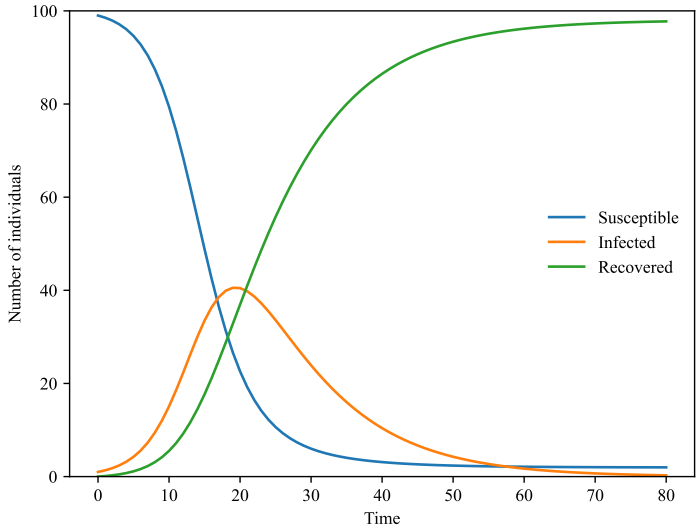

Figure 1.1: SIR model dynamics: Over 80 days, the number of susceptible, infected, and recovered individuals is tracked, illustrating the interaction dynamics.

stages. To capture this phenomenon, modelers move from deterministic mean-field models towards stochastic models, which allow for random events to occur. However, as differential equations are a powerful method for solving temporal state changes, this section showcases stochasticity in a differential equation model.

1.2.1 Stochastic Processes

Stochastic models use different *stochastic processes*, $\mathbf{X}$, to incorporate randomness. Now, a stochastic process is defined as a collection of random variables,

$$\mathbf{X} = \{X(t), t \ \in T\}, \tag{1.2}$$

where $X(t)$ represents the state of the stochastic process at time t in the time interval T. Time can either be discrete or continuous, depending on the type of process [81]. For a discrete-time stochastic process, we assign a random variable to each discrete time value. For example, random variables for $t \in \{0, 4\}$ could give the set $[0, 1, 0, 0, 2]$. On the other hand, continuous-time stochastic processes allows for independent time increments, meaning the state can change at any time on the interval $[0, t]$. In the context of disease dynamics, this set of random variables could indicate infected population counts, an infection event or events, along with a variation from the average of the system.

Different types of stochastic processes allow for additional properties to govern the process of interest. Stochastic models incorporate either one type or multiple sub-types of a stochastic process into their framework. For example, in an SI process, a stochastic differential equation (SDE) tracks the process $X(t) = [S(t), I(t)]$. SDEs are similar to compartmental models except for the noise added to the system through a Wiener process. The Wiener process is a stochastic process that provides change variable i.i.d. from a Normal distribution with $\mu = 0$ and $\sigma^2 = t$ [81, 5]. While the SDEs sample a single stochastic process over time, averaging those paths is an extra step to understand the probability a particular process occurring. To track a probability distribution with differential equations, we need a different model with a different stochastic process. The stochastic process focused on in this thesis is a Markov chain. The current state's dependence on only the previous state is defined as the *Markov property*, while successive events defined as such produce a *Markov chain* [81]. In particular, we focus on Markov chain's transition probabilities and a specific type of Markov chain, branching processes. Transition probabilities integrate

stochasticity into the model for Chapter 2, while branching processes are integral to the methods defined in Chapters 4 and 5.

1.2.2 Stochastic Simulation

Before over viewing the particular stochastic models presented in this thesis, we must address how stochastic models are validated. Ideally, real-world data of infection counts over time would for example be the source of validation for the models of this thesis. However, Chapter 2 details a cellular infection inside the body that, once found, is eradicated. This eliminates the possibility of data with progression information, leading to the need for stochastic simulation. Chapter 4 proposes intervention rollouts on an infected population, which are unique schemes and large amounts of data for those schemes are not available. Finally, Chapter 5 is performing perturbations or simulated noise on a system to evaluate the expected proportion of the population to be infected. Contact distributions with known noise and information on the proportion of the population infected is not available at this time. Therefore, we use simulations to validate our models.

For the Markov chain stochastic processes used in the Chapters 2 and 4 of this thesis, we perform simulations using a Gillespie algorithm [37]. This algorithm defines a continuous-time event-driven algorithm, meaning the algorithm determines the time at which *the next event* will happen. We note the perturbed systems simulated in Chapter 5 do not follow the algorithm in the next paragraph.

Simulations initialize an event-queue with a starting event, say an infected individual enters the population. Once this initial event is placed in the event-queue, it triggers the process to determine the time at which the next event will occur. A

Poisson process is used to model the time to the next event. This process defines the rates at which an event occur, also defined as the rate for a Poisson distribution. As a result, the time to the next event is given by an exponential distribution, which is parameterized by 1 over the event rate. A larger rate leads to a smaller exponential distribution parameter, meaning there is a higher chance of producing a smaller random time value to next event. Furthermore, for the smallest time, the corresponding event will be placed in the event-queue to occur at time equal to the current time plus the pulled time value. This algorithm comes to a stop when either there are no more events in the event-queue to occur, or the specified end time is reached.

1.2.3 Master Equations

Rather than track a single stochastic process, master equations track the full distribution of possible outcomes for these continuous-time stochastic dynamics. This stochastic process defines a fixed probability of the process moving from state i at time t to state j in time $t + \delta t$. This probability is defined as $\omega_t(i, j)$, which assumes that current state i, is the only state to inform the next state j. Remember, the chain of events for this stochastic process leading to $\omega_t(i, j)$ defines a Markov chain [81]. For master equations, the probability distribution of the system being in state i is $C_i(t)$ [39]. In this simple master equation example, i could indicate the number of infected cells. With probability flowing between states over time, modelers can track the probabilistic evolution with a differential equation structure. This is shown in Fig 1.2, where the transfer of probability occurs between neighboring states, with the resulting master equation for the system given by

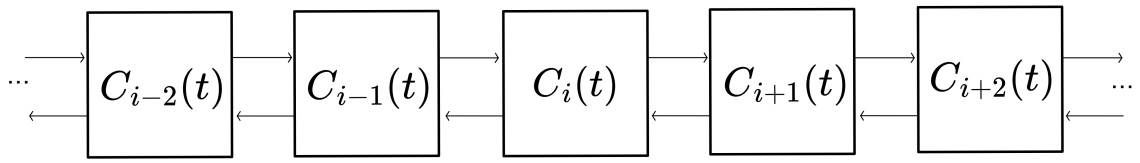

Figure 1.2: Probability flow for master equations: For the probability of being in state i, there is a transfer of probability to and from other states of the system. Each arrow defines an ω_t from one state to another.

$$\begin{aligned}\frac{d}{dt}C_i(t) = &-\omega_t(i, i-1)C_i(t) - \omega_t(i, i+1)C_i(t) \\ &+ \omega_t(i+1, i)C_{i+1}(t) + \omega_t(i-1, i)C_{i-1}(t). \end{aligned} \tag{1.3}$$

Since this equation defines the governing dynamics of $C_i(t)$, arrows leaving that state give negative terms. Similarly, arrows going towards $C_i(t)$ are represented as positive terms. The major advantage to tracking distributions lies in their moments, more importantly, how the moments can inform stochastic disease extinction and persistence information. More details and an application of master equations with the method of moments for with-in host disease modeling are shown in Chapter 2.

Chapter 2

Within-host infection dynamics with master equations and the method of moments: A case study of human papillomavirus in the epithelium

Abstract

Master equations provide researchers with the ability to track the distribution over possible states of a system. From these equations, we can summarize the temporal dynamics through a method of moments. These distributions and their moments capture the stochastic nature of a system, which is essential to study infectious dis-

eases. In this chapter, we define the states of the system to be the number of infected cells of a given type in the epithelium, the hollow organ tissue in the human body. Epithelium found in the cervix provides a location for viral infections to live and persist, such as human papillomavirus (HPV). HPV is a highly transmissible disease which most commonly affects biological females and has the potential to progress into cervical cancer. By defining a master equation model which tracks the infected cell layer dynamics, information on disease extinction, progression, and viral output can be derived from the method of moments. From this methodology and the outcomes we glean from it, we aim to inform differing states of HPV infected cells, and assess the effects of structural information for each outcome.

2.1 Introduction

Population models for the spread of disease are a major area of study which emphasize the macroscopic effect of a disease: the number of cases, complications, hospitalizations, or mortality within a population as a whole. As we refine these models, we also aim to capture the nuances of underlying within-host disease dynamics, which are often hidden under parameters and assumptions of population models. The intricacies of within-host interactions have been explored by many, as reviewed by Speranza [87]. Sequencing cells, identifying infection, tracking replication, and understanding the spatial aspect of cellular infection play into the complexities that can inform heterogeneous dynamics in a population [87]. From literature reviews of within-host biological dynamics, we attempt to understand the mechanisms of a disease. The focus of this chapter is on human papillomavirus (HPV), where the literature ad-

mits gaps in knowledge. Gravitt showcases HPV knowledge limitations in latency, clearance and incidence [38]. These limitations can lead to a single parameter that assumes homogeneity within the dynamics or calibrating the model parameters with data. This chapter takes a different approach. This work aims to use master equations and method of moments to illuminate the within-host cell dynamics of HPV that pose uncertainty in population models. [29, 82].

Modeling healthy and unhealthy cells in an epithelium is not a novel idea. Ordinary differential equations are used in many different ways to model within-host dynamics like epidermis dynamics [22]. Murall *et al.* focus on three systems of ordinary differential equations, two of which model skin systems for unvaccinated and vaccinated hosts exposed to HPV. The final system is a compartmental model for the transmission dynamics between individuals [69]. Both Sierra-Rojas *et al.* and Asih *et al.* define ordinary differential equations for the populations of differing layers of the epithelium [86, 8]. Another differential equation model for general epidermis turnover was developed by Ohno *et al.* [75]. Lastly, partial differential equations enter the space when Sari *et al.* consider how time and age affect the progression of HPV toward cancer [85]. While the mechanistic aspect of an epithelial infection is included in these models, they do not encapsulate the stochasticity that comes with infections.

All the aforementioned models are defined as mean-field models, which track the average dynamics of the system. Their structure makes it easy to define changes in average states of the system, however, the distribution of possible states around the average are missed. Mean-field models do not address the stochasticity that comes with infections. In the context of cell divisions, their random nature can cause

extinction events, resulting in transient infections. On the other hand, stochastic reinfection and multiple infection events can result in persistent infections. To factor in these types of infection events, models other than ordinary or partial differential equations have been explored.

Stochasticity not only integrates varying infection events, but provides perspective on the gaps in HPV knowledge [38, 82]. Branching processes are one stochastic method, which Ryser *et al.* and Beneteau *et al.* each use to address the randomness of cell division dynamics for HPV clearance [83, 12]. While modelers use branching processes to define the probability of random events, master equations strike the balance of tracking all aspects of a system with stochastic dynamics. Since probability distributions inherently provide stochasticity to a model, master equations track these distributions of all states in the system are tracked over time. This method accounts for the probability transfer between states, which is detailed in Sec. 2.2.2. One example of an infection-specific master equation approach is detailed by Vaughan *et al.*, who define target, HIV-infected, and virion cells, and the dynamics between them [96]. In contrast, Clayton *et al.* give an in-depth model using master equations to tracking a cell through its division process. The stochastic divisions of the base layer cells, either asymmetric and symmetric, maintain homeostasis in an arbitrary epithelium [23]. This chapter follows Clayton *et al.*'s model for the epithelial division process, however, their work focuses on the homeostasis achieved in epidermis tissue and general cell clone-size distributions [23]. Our results focus on how moments of distributions computationally simplifies the issue of solving a large master equation system. Avoiding this computational expense allows for additional cell types to be added in the system, meaning structural comparisons can occur. These moments

are defined through the method of moments, discussed in detail in Sec. 2.2.3, and allow for determining extinction, persistence, and viral output events for different structured systems.

Now, HPV is a sexually transmitted infection that is highly transmissible, leading to either transient or persistent infections. Transmission occurs through direct skin contact, causing infections to affect either the skin or mucosal epithelium. There are two categories of HPV infections: high and low-risk genotypes. Each type comes with different symptoms, for example, low-risk genotype infections lead to skin warts. High-risk genotype infections most commonly lead epithelial lesions, which can progress into cancer. These lesions and cancer are usually found in the cervix, but cancers of the vulva, vagina, penis, anus, mouth and throat are also possible. The Centers for Disease Control report that 99% of cervical precancers detect high-risk genotypes, and that one of these genotypes specifically cause about half of cervical cancers around the world [58].

Figure 2.1 depicts a cervical epithelium composed of *basal*, *parabasal*, *intermediate* and *superficial* cell layers. The basal cells mature through each layer to the superficial layer [70, 78]. In a healthy epithelium, basal and parabasal cells have the ability to divide, however, parabasal cells can also differentiate into intermediate cells. Once a cell is in the intermediate cell layer, it can only transition up to the superficial cell layer. Eventually, a superficial cell will shed and no longer be a part of the epithelium. An individual becomes infected with HPV when the virus infects the basal cell layer. The infection then propagates through cell divisions and transitions, but will only clear when there are no more infected basal cells [86].

With master equations providing the structure of differential equations and incor-

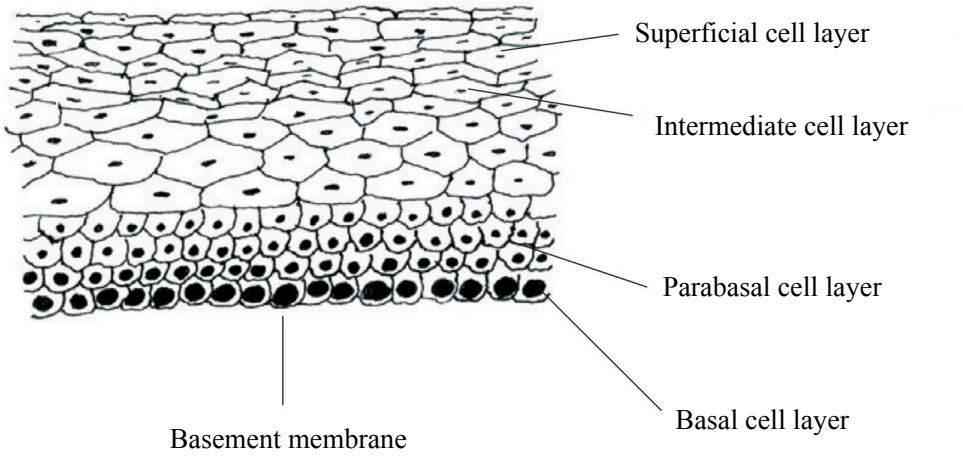

Figure 2.1: Cervical epithelium: Shown is the side view of the cervical epithelium, with its distinct cell layers. For the models presented in this chapter, the basement membrane will not be included. Cells either divide or differentiate, meaning they either duplicate or transition to the subsequent layer, to sustain each layers depth [23][1].

porating stochasticity into the system, we model HPV infection progression through epithelial tissue. We show the details of master equations and the method of moments in Sections 2.2.2, and 2.2.3. From this mathematical framework, we capture the mechanistic essence of extinction, persistence, and viral load in Sections 2.2.4, and 2.2.5. In this chapter, we simulate and derive analytical measures for two different systems in Sections 2.2.6, 2.3.1, and 2.3.2. Finally, Sec. 2.4 will validate the method of moments with simulations, along with showcasing the structural effect on each outcome.

2.2 Methods

2.2.1 Assumptions

In order to model this system mechanistically, we assert the following assumptions for HPV infections: First, our focus is only on specific divisions and cell processes which are described in Sections 2.3.1 and 2.3.2. Second, our model assumes there is only one infection occurring at a time, defined through initial conditions and lack of reinfection events, even though that might not be true for a real-world scenario and could be easily relaxed in our equations. Third, it is assumed that we are infecting the system with a high-risk genotype infection, to simulate infected cells that accumulate into lesions. Fourth, viral latency is an unknown aspect of HPV that Gravitt points out, however, this model does not incorporate a latent period in the spreading process [38]. Finally, we do not account for virion decay and instead focus on cumulative expulsion of virions from cells that have shed, or are considered dead. As stated previously, the assumptions of master equations are as follows: master equations define unique state changes that happen in continuous time. That being said, the number of such states is constrained by how computationally expensive it is to run the resulting system of equations. It is important to note that the master equations and method of moments are all exact, however, the extinction analysis is no longer exact since assumptions on the distributions are imposed. *The purpose of this model is to provide insight into epithelium dynamics modeling and the role of epithelium structure using the method of moments to reduce computational cost.*

2.2.2 Master equations

Consider the general first order linear differential equation, defining the instantaneous rate of change for the discrete variable y,

$$\frac{dy}{dt} + p(t)y = g(t). \tag{2.1}$$

The equations $p(t)$ and $g(t)$ define time-dependent processes [19]. When this framework is applied to population dynamics, where y represents the population density at time t, we can solve for an exact solution. Due to its deterministic nature, solutions give a specific *unique* population density at a time t. To transition from a deterministic framework to a stochastic framework, which accounts for the randomness of certain processes, while still using differential equation structure, we look to master equations.

In this stochastic framework, the discrete variable is replaced by the probability associated with the state y, denoted by $C_y(t)$. For all states, the probability distribution incorporates stochasticity, which is beneficial for integrating uncertainty into this model. Master equations also substitute equations $p(t)$ and $q(t)$ for transition rates, ω_t, which define a stochastic rate at which the state of the system changes. These stochastic rates can be thought of as the transfer of probability mass from one state to another [39]. We define a general master equation as,

$$\frac{d}{dt}C_y(t) = -\sum_z \omega_t(y,z)C_y(t) + \sum_z \omega_t(z,y)C_z(t), \tag{2.2}$$

which represents the change in the occupation probability or probability associated with state y. The left-hand summation of Eq. (2.2) defines the subtraction of mass from the probability density, $C_y(t)$, meaning the probability mass is leaving state y to any state z. With the addition of probability mass shown in the right-hand summation, we notice probability mass moving from any state z to y [39]. It is important to notice that the arbitrary state y represents all information relevant to the state, and can therefore be a scalar or a multi-dimensional quantity.

As an example, consider an arbitrary state which represents the scalar quantity n. This quantity tracks the number of infected cells in the system. The stochastic rates of transitions for this system are the infection rate, ω_i and the recovery rate, ω_r. We define three master equations for $n = \{1, 2, 3\}$ as

$$\frac{d}{dt}C_{n=1}(t) = -\omega_i C_{n=1}(t) - \omega_r C_{n=1}(t) + \omega_r C_{n=2}(t), \tag{2.3}$$

$$\frac{d}{dt}C_{n=2}(t) = -\omega_i C_{n=2}(t) - \omega_r C_{n=2}(t) + \omega_i C_{n=1}(t) + \omega_r C_{n=3}(t), \tag{2.4}$$

$$\frac{d}{dt}C_{n=3}(t) = -\omega_i C_{n=3}(t) - \omega_r C_{n=3}(t) + \omega_i C_{n=2}(t). \tag{2.5}$$

Now, we have a definition for the change in the probability density for $n = \{1, 2, 3\}$. From these equations we can determine the probability density at time t for each n. The total number of states for this system is the range of n, meaning we have three possible states to track and have two degrees of freedom. Since a single dimension master equation model represents the probability distribution dynamics for n, we can extend this framework to a multi-dimensional master equations. The multi-dimensional master equations have the ability to represent population interactions.

We define another quantity, h, representing the number of healthy cells. Therefore we can define $C_{n,h}(t)$ changing over time for $n = \{1, 2\}$ and $h = \{1, 2\}$ as

$$\frac{d}{dt}C_{n=1,h=1}(t) = -\omega_i C_{n=1,h=1}(t) - \omega_r C_{n=1,h=1}(t), \tag{2.6}$$
$$\frac{d}{dt}C_{n=2,h=1}(t) = -\omega_i C_{n=2,h=1}(t) - \omega_r C_{n=2,h=1}(t) + \omega_i C_{n=1,h=2}(t), \tag{2.7}$$
$$\frac{d}{dt}C_{n=1,h=2}(t) = -\omega_i C_{n=1,h=2}(t) - \omega_r C_{n=1,h=2}(t) + \omega_r C_{n=2,h=1}(t), \tag{2.8}$$
$$\frac{d}{dt}C_{n=2,h=2}(t) = -\omega_i C_{n=2,h=2}(t) - \omega_r C_{n=2,h=2}(t). \tag{2.9}$$

By adding another cell, we alter where the probability mass is transitioning from, which showcases the population interaction possible with this method. The number of states also increases to four, providing three degrees of freedom, which is achieved by multiplying each quantity range by the other. We will use this multi-dimensional approach moving forward, and specify the types of transitional interactions between each quantity in the system. As shown in the examples, this chapter defines a quantity as a unique cell type, therefore, the multi-dimensional system is renamed as a multi-cell-type system. Sections 2.3.1 and 2.3.2 define a three-cell-type system, representing a three layer epithelium, and a five-cell-type system, representing a five layer epithelium.

2.2.3 Method of moments (MoM)

From the small multi-cell-type example in Sec. 2.2.2, we define a system with two cell types and see the number of equations to satisfy with a solution grows exponentially

with the number of cells. For multi-cell-type models with large ranges, the state spaces grows quite large and this explicit method becomes computationally expensive. A simple solution to avoid solving for a large number of states is to derive the mean and variance from the probability distribution being defined. A mean value and variance description has been deemed adequate for describing large state spaces or populations [39]. This process is not as computationally expensive as explicitly solving all the master equations, since there are only the chosen moments of each state and their interaction terms to track over time.

The mean for a given cell, k, in the cell range of 0 to L, $y = (y_1, y_2, y_3, \ldots, y_L)$ is,

$$\bar{y}_k(t) = \sum_y y_k C_y(t) = \langle y_k \rangle_t, \tag{2.10}$$

where $C_y(t)$ is the probability distribution for all states. The master equations solve for the change over these probability distributions, so using Eq. (2.10), we define the equation for the change in the moment of a distribution as

$$\frac{d}{dt}\langle y_k \rangle_t = \sum_y y_k \frac{dC_y(t)}{dt}. \tag{2.11}$$

Now, given a multi-cell-type system, we must consider higher order moments and interaction terms as

$$\frac{d}{dt}\langle y_k^\ell y_s^q \rangle_t = \sum_y y_k^\ell y_s^q \frac{dC_y(t)}{dt}. \tag{2.12}$$

From the first and second moment of state y, we define the variance as

$$Var(y_k)_t = \langle y_k^2 \rangle_t - \langle y_k \rangle_t^2. \tag{2.13}$$

2.2.4 Extinction and non-extinction events

The probability of extinction is essential for understanding the duration of an infection event, therefore, from the first two moments of any cell, we can derive the probability of extinction. We can do this from the assumption that the underlying distribution of a given average, $\langle y_k \rangle$, follows a zero-inflated geometric distribution. Kendall previously showed that the geometric distribution is a solution to the birth-death process [47] therefore, we use a geometric approximation of non-extinct trajectories to extract the probability of extinction from the method of moments. We redefine the first two moments as,

$$\langle y \rangle = \frac{1}{p}[1 - P(y = 0)], \tag{2.14}$$

$$\langle y^2 \rangle = (0)P(y = 0) + \frac{2 - p}{p^2}[1 - P(y = 0)], \tag{2.15}$$

where $P(y = 0)$ is the probability of extinction for cell y and $\frac{1}{p}$ and $\frac{2-p}{p^2}$ are the first and second moment of the geometric distribution for non-extinct states. Solving for p from Eq. (2.14), we define

$$p = \frac{[1 - P(y = 0)]}{\langle y \rangle}. \tag{2.16}$$

Now, substituting Eq. (2.16) into Eq. (2.15), we can solve for the probability of extinction as

$$P(y = 0) = 1 - \frac{2\langle y \rangle^2}{\langle y^2 \rangle + \langle y \rangle}, \tag{2.17}$$

which provides a solution for the probability of extinction using the first two moments of the distribution of associated with cell y. In Sec. 2.4, details on transient infections defined by the probability of extinction will be discussed. Now, when solving for the probability of extinction, we inherently solve for the probability of non-extinction as well,

$$1 - P(y = 0) = \frac{2\langle y \rangle^2}{\langle y^2 \rangle + \langle y \rangle}. \tag{2.18}$$

From the probability of non-extinction, we can estimate the geometric distribution parameter with a specific cell from Eq. (2.16),

$$p = \frac{2\langle y \rangle}{\langle y^2 \rangle + \langle y \rangle}, \tag{2.19}$$

thus estimating the mean and variance of a cell count for a persistent infection. This provides valuable insights for the average basal cells from a persistent infection, which will be discussed in Sec. 2.4.

2.2.5 Viral load

The viral load metric in an individual is the product of infected cells shedding from the epithelium. When a cell is shed, it disperses virions or copies of HPV that can go on to infect others. As pointed out in Sec. 2.1, there are many uncertainties around the progression of an HPV infection, one of which is viral load [38, 82]. It is known that when a cell becomes infected by HPV, there is a genome replication that occurs in the basal cell layer, resulting in 50-100 viral copies in the cell [26, 88]. As infected cells differentiate and move up the cell layers, more viral copies are produced within the cell. The number of copies varies according to the genotype, however, we set the

number of virions to 1,000 as a proof of concept. We determine viral load output from the MoM by focusing on the moments associated with the final cell type, dead cells. We determine the first and second moment of the dead cells, then apply the virion shed value for each dead cell, producing the expected total number of virions released over time.

In the future we can aim to give a general distribution for the viral copies per cell [33, 94]. When a cell is shed from the epithelium, it is estimated between $50-10^4$ viral copies are expelled. [88, 66, 33, 94]. Due to this wide range for viral copies shed out of the system, we can apply a specified distribution for the previously mentioned range. This distribution could be altered according to new research or other assumptions, for example, to include more stochasticity, the distribution over the given range could be uniform.

2.2.6 Stochastic simulation algorithm

Without data for these cellular dynamics, we simulate the stochastic process of an infection in the cervical epithelium with the Gillespie stochastic simulation algorithm [36]. As described in Sec. 1.2.2, this algorithm establishes a continuous-time simulation, which tracks the events that occur as time passes. These event-driven simulations are conducted by distinguishing the rates of each event, then assigning each rate to its respective exponential distribution. Drawing from all the exponential distributions provides the next time to all specific events. Whichever time is the closest to the current time means the associated event will occur. New events based off the event that just occurred are placed in the queue, since one event triggers a future event. This continues until time runs out or all cell types are equal to 0, excluding

the dead cells. For the purposes of this project, we set a maximum time of 750 days to occur for 50,000 simulations.

2.3 Application to structured epithelium dynamics

2.3.1 Three-cell-type system

For the first iteration of this model, we will only focus on the first two layers of the cervical epithelium, the basal and parabasal cell layers. The biological reason for focusing on the bottom two layers first is that after menopause, the cells in the cervix do not mature past the parabasal cell layer. This results in a thinner cervical epithelium [78]. After an initial basal cell is infected, set as an initial condition, there are three types of basal cell divisions that can accumulate more infected cells in the epithelium. An infected basal cell can divide into two more infected basal cells, divide into an infected basal cell and an infected parabasal cell, or divide into two infected parabasal cells. These divisions happen with rates of β, γ, and δ respectively. Once an infected parabasal cell enters the system, it has the potential to divide into two more infected parabasal cells, or differentiate, which allows the cell to shed from the epithelium. Each of these processes occur with respective rates of ρ and θ. The resulting system is therefore fully specified by the number of infected basal cells, b, the number of infected parabasal cells, p, and the number of infected dead cells shed, d. The epithelial dynamics between these three types of cells are depicted in Fig. 2.2. The parameter values are shown in Table 2.1, which are used in the MoM equations

and simulations.

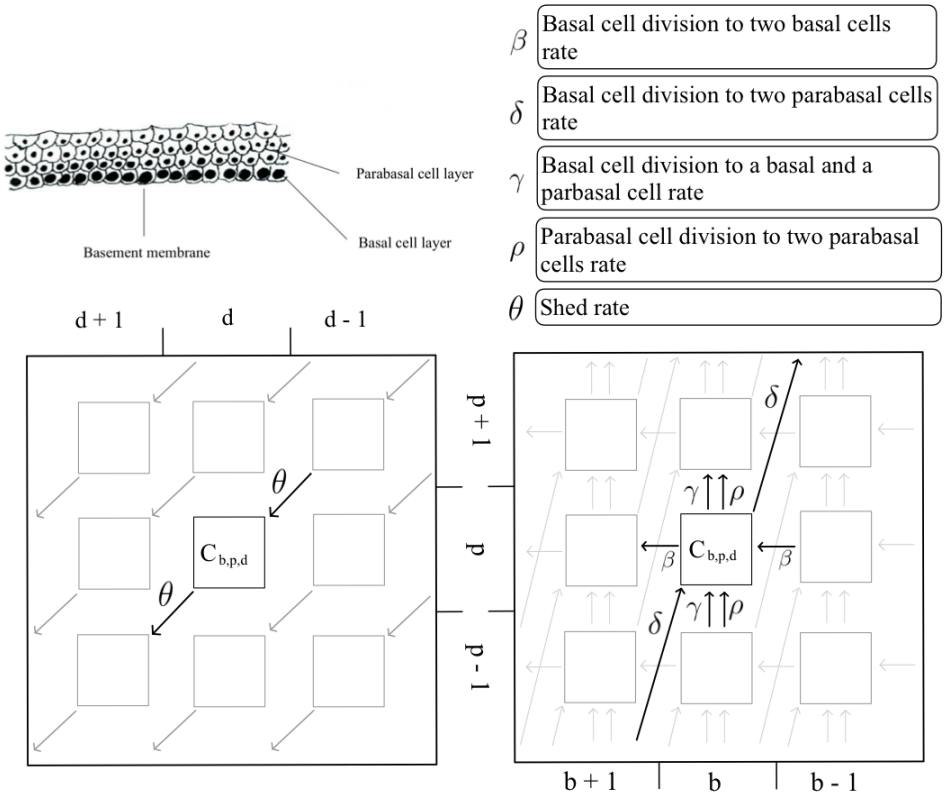

Figure 2.2: Three-cell-type model schematic: Each state of this system is fully specified by the number of infected basal cell b, the number of infected parabasal cells p, and the number of infected dead cells shed d. The general dynamics are shown with arrows illustrating the probability mass transitions between states. A state change is made when the infected basal cell count changes, the infected parabasal cell count changes, or the dead cell count changes in the system.

The master equation derived from Fig. 2.2[2] tracks three quantities: b, p, and d. The general probability distribution over all possible states $\{b, p, d\}$ is defined as $C_{b,p,d}(t)$, which evolves in time according to,

[2]Cell image reprinted from Colposcopy and Treatment of Cervical Precancer, IARC technical publication No. 45, Walter Prendiville and Rengaswamy Sankaranarayanan, Anatomy of the uterine cervix and the transformation zone, Page 14, 2017 (2017) [78].

$$\frac{d}{dt}C_{b,p,d}(t) =(b-1)\beta C_{b-1,p,d} + (p+1)\theta C_{b,p+1,d-1} + ((p-1)\rho + b\gamma)C_{b,p-1,d} + (b+1)\delta C_{b+1,p-2,d} - [b\beta + b\gamma + b\delta + p\rho + p\theta]C_{b,p,d}. \quad (2.20)$$

As explained previously, if we track our counting variables b, p and d up to some large integer N, we are then dealing with a system of N^3 equations which can become unwieldy when dealing with realistic infection sizes. Thus we move to the method of moments as a computationally inexpensive alternative. This work focuses on the first and second moments of all the cell variables. The exact equations are defined in the Supplemental Materials. The derivations of these moments follow from Eq. (2.11). Numerical solutions to these equations give enough information to derive the distribution's mean and standard deviation for all cell-types at time t. These exact analytical results can the be compared to the outcome of the simulation process defined in Sec. 2.2.6 starting from a single infected basal cell. The simulation then tracks the history of each cell type count over the 750 days.

	Process definition	Rate [1/days]	Reference
β	Basal cell to two basal cells division	0.0034	[12, 70, 23]
δ	Basal cell to two parabasal cells division	0.0024	[12, 70, 23]
γ	Basal cell to one basal cell and one parabasal cell division	0.0252	[12, 70, 23]
ρ	Parabasal cell to two parabasal cells division	0.0312	[70, 23]
θ	Parabasal cell shed	0.67	[70]

Table 2.1: Three-cell-type system parameters: Parameter values for the processes occurring on the system are used in both the analytical model and the simulations for the three-cell-type system. All rates are measured in units of 1/days.

2.3.2 Five-cell-type system

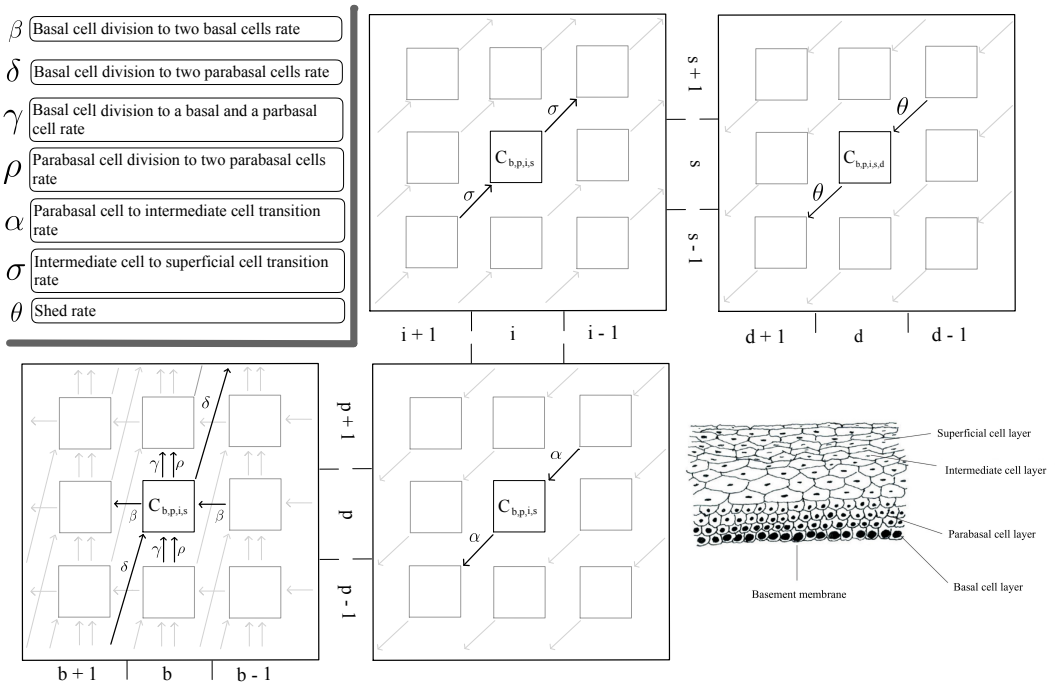

Figure 2.3: Five-cell-type model schematic: Similar to the three-cell-type system, the general dynamics for the five-cell-type system are shown with arrows illustrating the probability mass transitions between states. A state change is made when any of the infected cell counts or the dead cell count change in the system. With the additional cell types, we see the differing dynamics from parabasal cells to intermediate and superficial cells.

The second iteration of this model adds the dynamics of the two upper layers in the cervical epithelium, the intermediate, i, and superficial, s, cell layers. The general probability distribution of infected cells for each cell type is changed to $C_{b,p,i,s,d}(t)$ to incorporate the extra layers. Similar to Sec. 2.3.1, once an initial basal cell is infected, infected cells are introduced to other layers through divisions or transitions. Previously, we also described how infected basal cells can divide into two infected basal cells, an infected basal cell and parabasal cell, or two infected parabasal cells. Remember, these divisions occur at rates of β, γ, and δ respectively. Now, the intermediate and

superficial cells do not result from division dynamics, rather, they populate through transition dynamics. Therefore, the dynamics of the infected parabasal cells diverge from the three-cell-type model dynamics. Instead, an infected parabasal cell either divides into two infected parabasal cells with rate ρ, or transitions into an infected intermediate cell with rate α. Furthermore, intermediate cells cannot divide, so, the only event that can occur is for the cell to transition up to an infected superficial cell with rate σ. Lastly, an infected superficial cell will die and shed any virions with rate θ [78]. The interactions between subsequent layers are shown in Fig. 2.3[3]. The parameter values for each process rate are defined in Table 2.2.

The master equation derived from Fig. 2.3 details the flow in and out of each state. The equation is as follows,

$$\begin{aligned}\frac{d}{dt}C_{b,p,i,s,d}(t) =&(b-1)\beta C_{b-1,p,i,s,d} + ((p-1)\rho + b\gamma)C_{b,p-1,i,s,d} \\ &+ (b+1)\delta C_{b+1,p-2,i,s,d} + (p+1)\alpha C_{b,p+1,i-1,s,d} + (i+1)\sigma C_{b,p,i+1,s-1,d} \\ &+ (s+1)\theta C_{b,p,i,s+1,d-1} - (b\beta + p\rho + b\gamma + b\delta + p\alpha + i\sigma + s\theta)C_{b,p,i,s,d}.\end{aligned} \tag{2.21}$$

Similar to the three-cell-type system, we derive the first and second moments for each of the five cell types, along with the pairwise interaction terms. All exact MoM equations for the five-cell-type system are given below. The five-cell-type simulation follows the same process as described in Sec. 2.3.1, except the i and s states are also tracked. The exponential distributions for the time to next event considers the additional processes of parabasal cells moving to the intermediate cell layer, interme-

[3]Cell image reprinted from Colposcopy and Treatment of Cervical Precancer, IARC technical publication No. 45, Walter Prendiville and Rengaswamy Sankaranarayanan, Anatomy of the uterine cervix and the transformation zone, Page 14, 2017 (2017) [78]

	Process definition	Rate [1/days]	Reference
β	Basal cell to two basal cells division	0.0034	[12, 70, 23]
δ	Basal cell to two parabasal cells division	0.0024	[12, 70, 23]
γ	Basal cell to one basal cell and one parabasal cell division	0.0252	[12, 70, 23]
ρ	Parabasal cell to two parabasal cells division	0.0312	[70, 23]
α	Parabasal cell differentiating and moving to the intermediate cell layer	0.4	[70]
σ	Cell moving from the intermediate cell layer to the superficial cell layer	0.4	[70]
θ	Superficial cell shed	0.67	[70]

Table 2.2: Five-cell-type system parameters: Parameter values used in both the analytical model and the simulations for the five-cell-type system. All rates are measured in units of 1/days.

diate cells moving to the superficial cell layer. Finally, superficial cells shed out of the epithelium and add to the cumulative number of dead cells.

2.4 Results

Section 2.3.1 defines the analytical moments and the event-driven simulations for the three-cell-type system over time. Figure 2.4 shows the results and exactitude of the MoM equations compared to the simulations. As time goes on, the average number of infected basal cells grows slowly to 2 by 750 days. This indicates that the basal cell division rate is not large enough to double the average until after 750 days. The variance of the basal cells gradually increases over time as well, illustrating that the longer the infection persists, the more variable the basal cell counts can be. Turning to the average count of infected parabasal cells, the average jumps from 0, implying immediate divisions of infected basal cells to infected parabasal cells. The change in the infected parabasal cell average is not as large as the change in the infected basal

cell average, but intuitively we attribute this to the rate of shedding being larger than all other rates in the system. It is noteworthy that the simulated average of infected parabasal cells is more variable over time, which is a result of the small values for the average infected parabasal cells. This scale could also have an effect on the variance of the parabasal cells, which is small relative to the other cell-type variances. Moving on to the dead cells, there is a steady increase in the average over the time period, similar to the average basal cells. However, the variance of the dead cells grows at a fast rate, likely due to the large shed rate and the variability from the average parabasal cell counts.

With the intermediate and superficial cell layers to consider, the same analyses conducted for the three-cell-type system are conducted for the five-cell-type system. Figure 2.5 shows that with two more cell types, the simulations still validate the MoM results. The figure provides similar insights as Fig. 2.4 for the basal cell average and variance. On the other hand, the average and variance for the parabasal cells exhibit larger quantities over time. Turning to the average count of intermediate and superficial cells, both exhibit a similar trend as the average parabasal cell count. One major difference is the average count for parabasal and intermediate cells are larger than the average count for superficial cells. While the magnitude of the variances differ for the parabasal, intermediate, and superficial cells, each have a slow growing variance. This slow growth and variability in simulations is a result of the small averages of each cell type. Sec 2.2.5 compares the dead cell averages in terms of viral load to highlight the differences between the two systems.

By accurately modeling the mean and variances of each cell type, we have the ability to understand the distribution of cell type counts over time. Furthermore,

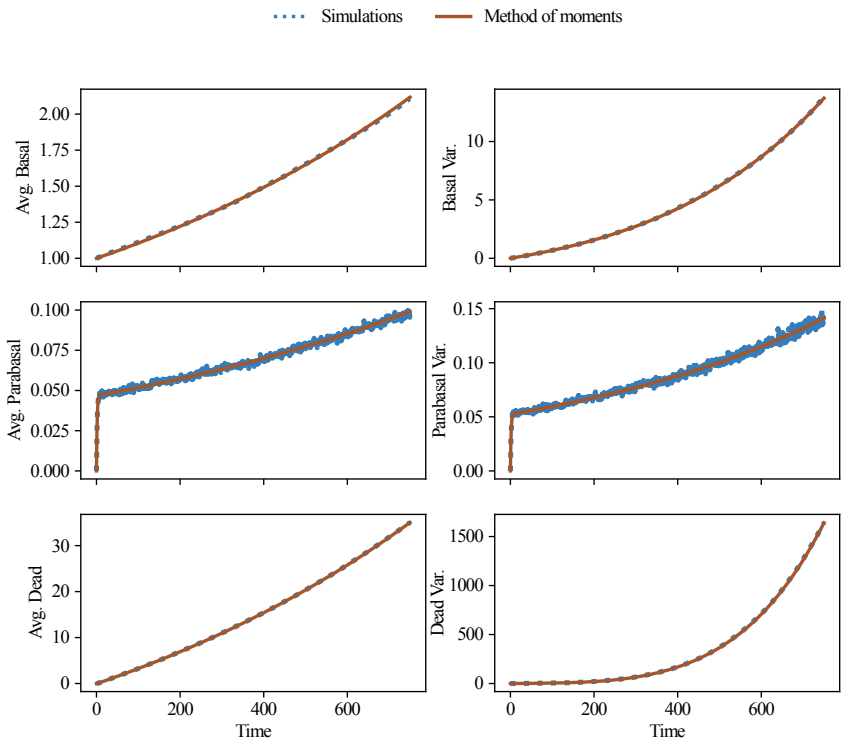

Figure 2.4: Average and variance evolution of the three-cell-type system: The averages and variances from the MoM equations for the basal, b, parabasal, p, and dead, d, cells over 750 days (solid lines) are validated by the simulations (dotted lines). 50,000 simulations were used to compute the mean and variance for each cell type. The initial conditions for the MoM and simulations set the first and second moments of b equal to 1, while the other moments are set to 0.

these distributions can encapsulate the likelihood of various disease progressions, for example viral load output, shown in Section 2.2.5. With the MoM being exact for all cell types, we assert that this method has the potential to derive probabilities of extinction, persistent infection indicators and average cumulative virion counts to inform infectiousness of an individual in population models.

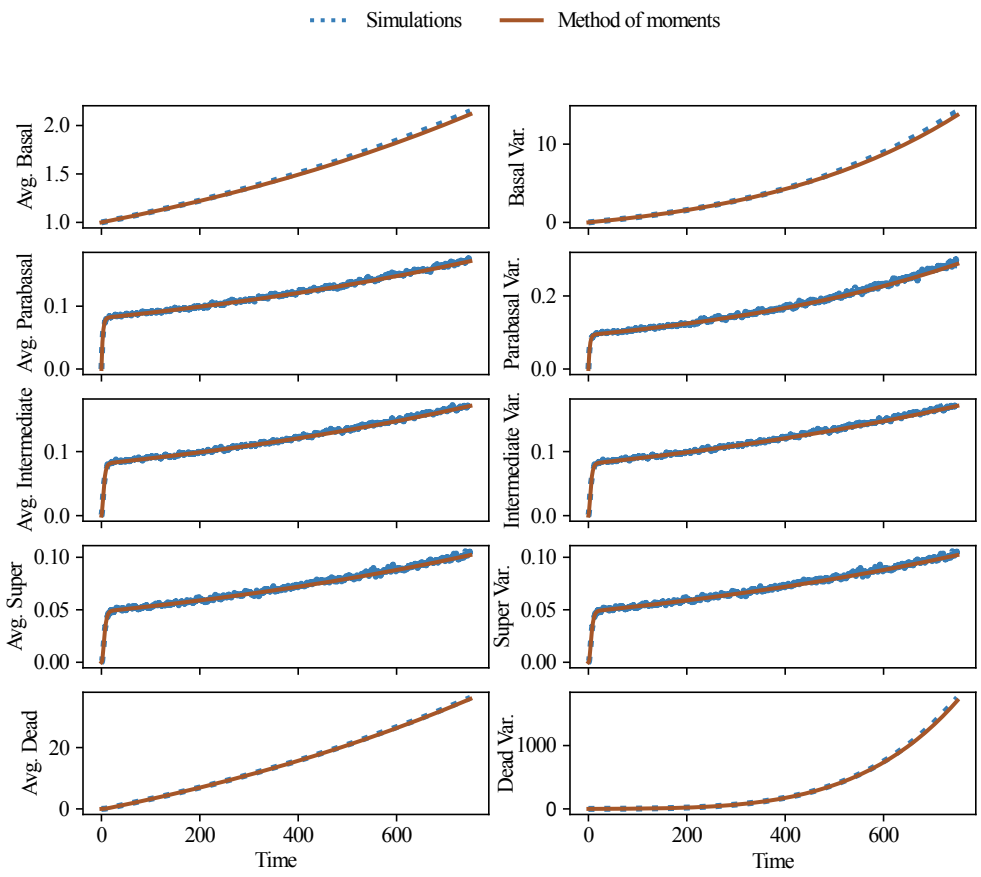

Figure 2.5: Average and variance evolution of the five-cell-type system: From the first and second moments of the basal, b, parabasal, p, intermediate, i, superficial, s, and dead, d, cells over 750 days, we derive the mean and variance (solid lines) for the five-cell-type system. The initial conditions are the same as the three-cell-type system, where the first and second moments for the basal cells are set to 1. All other moments are set to 0 at the start. We validate these analytical results with simulations (dotted lines). Similar to the three-cell-type system, 50,000 simulations were used for this validation.

2.4.1 Extinction probability

From the biological mechanisms of the epithelium, once the infected basal cell value is equal to 0, the parabasal cells will divide and eventually shed, eradicating the infection. Therefore, when the infected basal cell count reaches the zero-th column

of Fig. 2.2 and 2.3, this marks an extinction event. As shown in Sec. 2.2.4, the probability of the infected basal cell average equaling 0, from our assumed zero-inflated geometric distribution, is equal to the probability of extinction. Remember this is not an exact result since we cannot assume a geometric approximation to be exact for all times. The analytical derivation is validated with simulations in panel A of Fig. 2.6 for both modeled systems. We note that the differing systems do not affect the results of the extinction probabilities due to consistent basal cell dynamics, as shown in panels A and B. From the efficient MoM results, we are able to provide extinction probabilities for varying time periods. For example, panel A illustrates the probability of the average basal cells going extinct is over 50% at 750 days. Extending past the 750 day mark in panel B, the probability of extinction levels out around 70%. This means, after 6,000 days, there is a 70% chance the infection has cleared out of the basal layer. While the likelihood of clearing an infection is high, there is still a chance of persistence, which is discussed in Sec. 2.4.2.

2.4.2 Persistent infections

The dynamics of persistent infections are necessary for understanding the long term effects of an HPV infection. Therefore, we show the average basal cells without extinctions in Fig. 2.7 for both the three and five-cell-type system. Simulations excluding extinction events validate the MoM results for the average number of basal cells. Comparing the average basal cells in Figs. 2.4 and 2.5, the average basal cells for non-extinct infections are much larger. Consequently, the larger this average is over time, the longer the infection could persist past this point. Moreover, these findings estimate the severity of 30% of HPV infections that will persist over time.

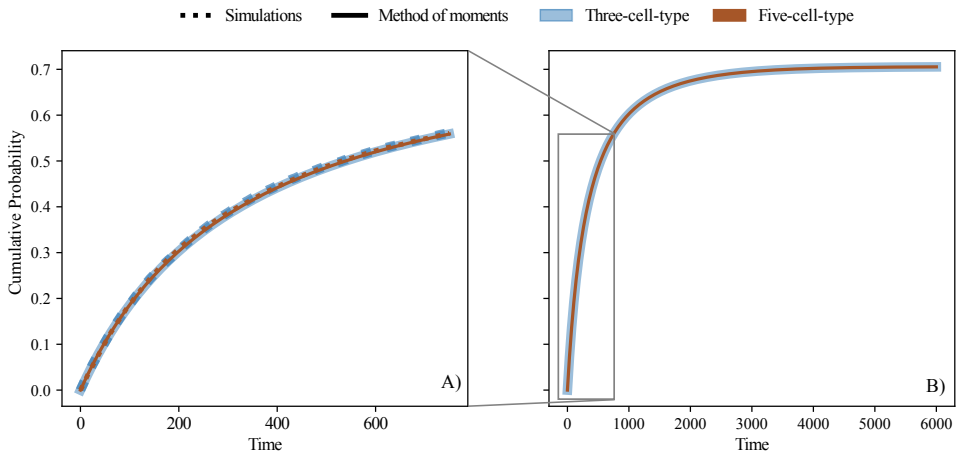

Figure 2.6: Cumulative probability of extinction of the basal cells for the three and five-cell-type system: In panel A, 50,000 simulations (dotted lines), over a 750 day time period, validate the probability of extinction derived from the first and second moments (solid lines) of the basal cells for the three (blue) and five-cell-type (red) system respectively. To derive the probability of extinction, we assume a zero-inflated geometric distribution, where we define the probability of extinction in terms of the first two moments. Panel B exhibits the probability of extinction after 6,000 days. Evaluating the probability of extinction past 750 days showcases the probability plateauing to 0.7 after 3,000 days.

Persistent infections not only affects the infectiousness of an individual, but also their risk for cancer. The progression to cancer and infectiousness play vital roles in the disease spread and mortality rate in population-level models.

2.4.3 Cumulative virions

Whether an infection is transient or persistent, we establish the average cumulative virions resulting from the epithelial dynamics. By assuming a certain number of virions shed from each dead cell, we achieve Fig. 2.8. This figure validates the MoM with simulations after a constant multiplier is applied to the average dead cell count for the three and five-cell-type system. The average cumulative virion count steadily grows as time goes on, ending at a little over 35,000 virions expelled at 750 day mark

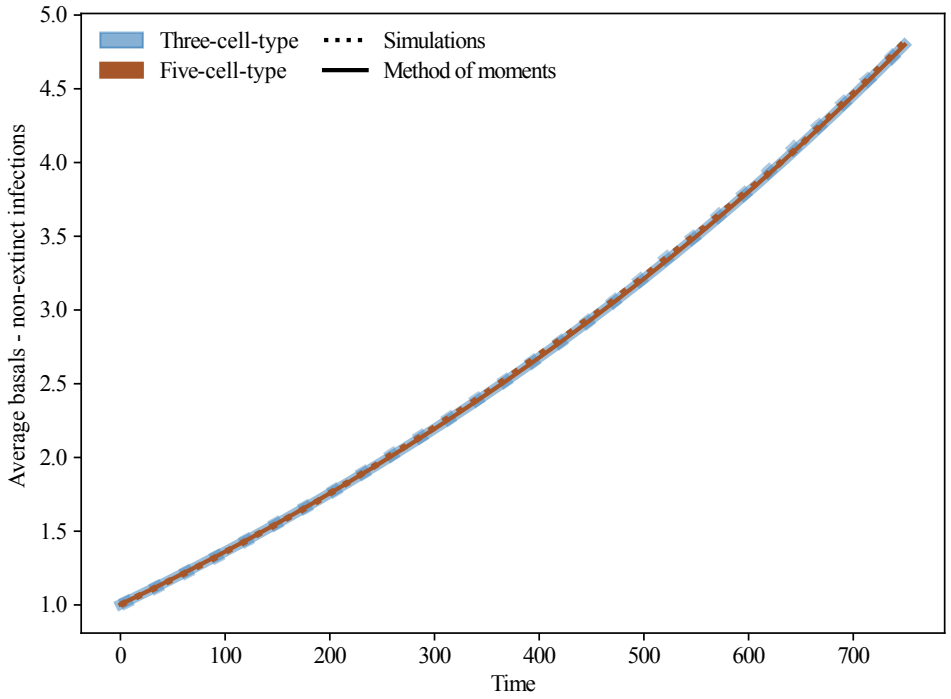

Figure 2.7: Average basal cells of persistent infections for the three and five-cell-type system: The non-extinct average basal cell simulation results (dotted lines) validate the MoM results (solid lines). The averages grow faster than those that include extinction events for both the three (blue) and five-cell-type (red) systems.

for the five-cell-type system. The two systems begin to significantly diverge between 150-250 days. Furthermore, the inset plot shows the difference between the three and five-cell-type system over time (pink solid line). For comparison, a quadratic fit is also plotted in the inset (black dashed line), showing the difference follows a quadratic trend for the 750 day period. As time passes, the larger the gap is between the two systems, exhibiting the important role the extra layers have in harboring more infected cells. Since the importance of the extra layers depend on a person's age [78], our results suggest that different approximations for an individual's level of infectiousness over time should be used based on their age to better inform population

models.

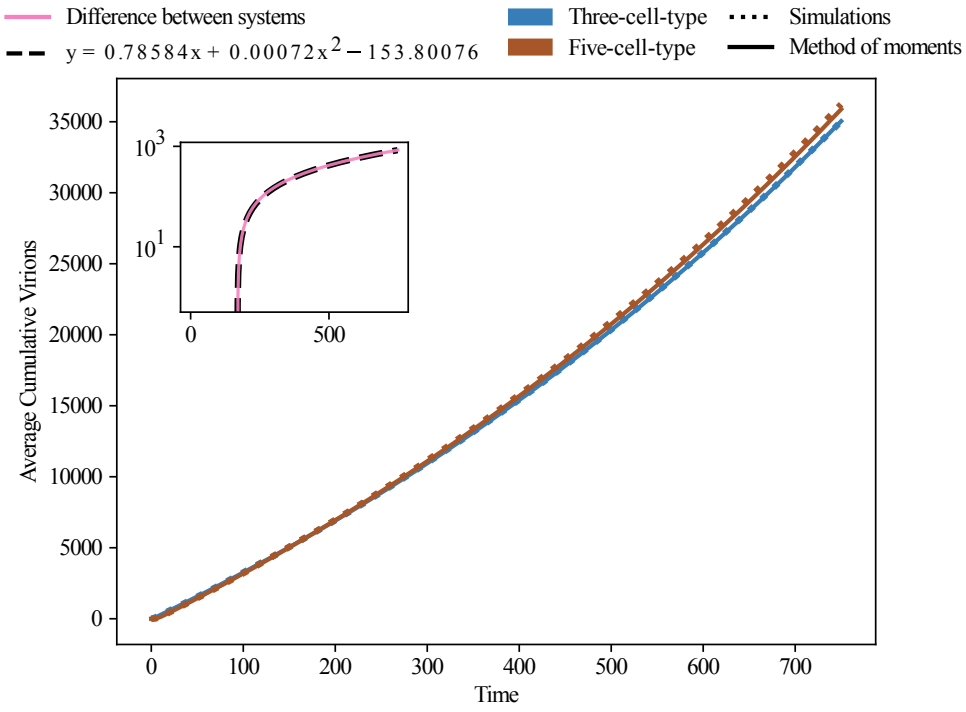

Figure 2.8: Average cumulative virions shed for the three and five-cell-type system: The average cumulative virion count from simulations (dotted lines) validate the MoM results (solid lines). The average cumulative virions shed is defined by 1,000 virions shed per dead cell. The difference in average cumulative virion count for the three (blue) and five-cell-type (red) system is highlighted in the inset plot (pink solid line), with a quadratic fit (black dashed line).

2.5 Discussion

While population level HPV models help inform persistent disease spread, within-host models have the ability to add individual-level stochasticity and heterogeneity to those population models. This is especially true since the mechanisms of cell division by which HPV infections propagate through an epithelium are well known. However,

other less known factors such as clearance and role of epithelium layers must also be considered. With the uncertainty in the outcomes of an HPV infection preventing accurate modeling, it is imperative to leverage the knowledge of cell division dynamics to infer important infection dynamics.

In this chapter, we have demonstrated how master equations can track the probability distributions associated with count of infected cells in a cell layer. For each system, we define a generalized governing master equation for the probability of being in an arbitrary state. While the solutions from these master equations are exact, the larger the system, the more computationally expensive it is to solve. To avoid this, we can instead simply track the statistical moments of the full state distribution using the method of moments. Rather than tracking a full distribution for every cell type over time, these moments provide an accurate and efficient way to study the probabilistic nature of HPV infection progression. Moreover, we can still use the derived moments to compute the probability of extinction, average basal cell count for persistent infections, and average cumulative virions over time. In doing so, we tested the effects of adding structure to the epithelium in the form of extra cell layers whose importance are known to depend on age. We found that while this extra structure does not affect the probability of establishment (or extinction) of a new infection, it does affect long-term shedding rates of virions in persistent infections.

Some considerations for future work are incorporating reinfections, and within-host spatial tracking. For simultaneous infections, it would be important to allow current infections to reinfect the host stochastically, or introduce another transmission event to add to the viral load burden. This consideration could result in a change to the moments of each cell-type distribution, meaning the probability of extinction

could go down drastically depending on the time of the reinfection event while the non-extinct average basal cell count could increase. This compounding effect could spike the average cumulative virions and change the level of infectiousness for the individual. Finally, incorporating a spatial component to this model could account for simultaneous infections merging to form a larger infected area. Consequently, this would account for possible lesion grades and treatments to rid an individual of the infection.

While this framework informs HPV infection growth, epithelium infection progressions do not only show up in a cervical epithelium. This model and its outcomes can be applied to other parts of the body as well. Leveraging this framework for not only other parts of the body but for other diseases could benefit modelers that struggle with individual-level disease knowledge. The generalisability of master equations and the method of moments for infection progressions through cell division dynamics provide a powerful framework that can aid population-level models.

The reasons to model span from understanding population-level disease dynamics to mitigating outbreaks, but there are large assumptions placed on modeling diseases that have a number of unknown aspects. Using the information on the within-host mechanisms of the disease creates an opportunity to build model pipelines. These pipelines would be composed of mechanistic and stochastic within-host models of diseases, which outcomes can inform parameters for population-level models. This informed heterogeneity can alleviate the pressure of calibration or unknown parameter values in large models by leveraging known mechanisms of disease propagation in the body. When struggling with the unknown knowledge of disease progression and transmission, leaning on known biological mechanisms can provide insights to fill our

gaps in knowledge.

Chapter 3

Branching Process Models of Disease Dynamics

3.1 Branching processes of disease spread

Master equations define the transitions between states and track the probability of being in a state over time, which is conducive for a system that has a homogeneous structure and minimal discrete states. However, with population disease spread, the evolution of infected individuals through heterogeneous structures is essential to capture the effect of disease spread on complex social structures. To allow for heterogeneity, we initiate a Markov chain on a structure with defined heterogeneity. This particular stochastic process is known as a branching process, which focuses on the generational cumulative count for a specified process, along with evaluating extinction events.

The inception of branching process analysis started in 1874 with Francis Galton and H.W. Watson considering the issue with family names going extinct over time

[42, 6]. This process considers an object in the zero-th generation that will create or 'branch' to another object or objects in the first generation. As detailed in Sec. 1.2.1, because the current generation is dependent on the previous generation, branching processes are Markovian processes.

Now, branching processes can result in extinction outcomes and cumulative counts in discrete time, which are important aspects to consider for the spread of a disease [40, 10, 11, 6]. The extinctions and cumulative counts are tracked through observations of the transmission chain or branch of disease spread. As this chain continues to grow and accumulate infected individuals, this describes the infection percolating in the population structure. Percolation defines the macroscopic effect of the infection spread on a structured population [89]. Moving forward, the focus will be on percolation on contact networks.

3.2 Percolation on Contact Networks

3.2.1 Network Definitions

The concepts of network science originate from graph theory. Graph theory itself roots from Leonhard Euler's 1741 work on the Könisberg bridge problem: a combinatorics problem of crossing seven bridge only once given the city structure of Könisberg [30]. Outside of city strutures, networks help understand the structure of power grids, the internet, social connections, and many more [71].

A network is defined as a collection of *vertices or nodes* that are connected by *edges*. Figure 3.1 provides an example of a small random network. A vertex can

represent an individual or one entity of the network. An edge then represents an interaction, or relationship between two vertices. When there is an edge between two vertices, those two vertices are neighbors of each other. The number of neighbors defines the *degree* of the vertex. In network epidemiology, vertices represent individuals or objects that could be infected and the edges between them are the avenues of transmission. This is defined as a *contact network*. One way to describe a contact network is by its *degree distribution*. This distribution defines a probability distribution for a vertex having a certain degree. We can define a degree distribution from an existing network, by counting the number vertices with each distinct degree then dividing by the total number of vertices to compute their frequencies. If the exact network structure is not given, a general probability distribution can define the degree distribution, providing a random graph or network [35]. For this thesis, we govern our contact networks with probability distributions. While any distribution can describe a contact network, epidemiologists focus on negative binomial distributions [34]. A negative binomial distribution can be parameterized to define R_0, the average number of secondary infections, and either α or k, a dispersion parameter defining how homogeneous or heterogeneous the distribution is.

3.2.2 Percolation on Contact Networks

As individuals on a contact network become infected through branching processes, *percolation* can occur. Percolation, a concept initially used to describe how molecules connect together to make macro-molecules, describes a branching process permeating through a structure, for this work, through a contact network [89]. There are two types of percolation that can occur on a network. The first being when all vertices

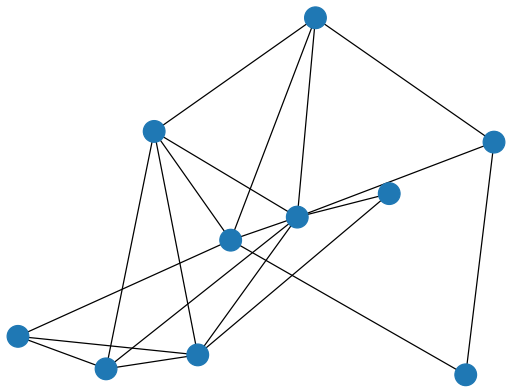

Figure 3.1: Basic random network: This figure illustrates 10 vertices randomly connected according to a uniform distribution by 20 edges, creating a random network.

have a probability defining their susceptibility. Similar to how a person may come in contact with a disease but their immune system may or may not fight it off. This is referred to as site percolation. The second, known as bond percolation, assumes there is a probability of transmission along the edges of the network. If no assumption is placed on immunity for the system, which we will see in Chapter 4, then with probability T, an infected vertex will infect one of their contacts over the course of the epidemic [73, 62, 72, 59]. While the networks in this paper only impose percolation on a random network, others have imposed more constraints on the networks they percolate on [63, 60, 67].

To define the probability of transmission, T, this work, along with others, assume transmission and recovery follow simple Poisson processes occurring at fixed rates β and γ respectively [43, 44]. It is assumed a vertex, when infected, is infectious for some random time τ. Therefore, the probability of the vertex infecting one of its neighbors during τ is

$$T(\tau) = 1 - \lim_{\delta t \to 0}(1 - \beta\delta t)^{\tau/\delta t}$$
$$= (1 - \exp^{-\tau\beta}). \tag{3.1}$$

In a similar manner, the probability of the infectious period being length τ is given by the cumulative distribution function over τ evaluated for the average rate of recovery, γ,

$$F(\tau) = 1 - \lim_{\delta t \to 0}(1 - \gamma\delta t)^{\tau/\delta t}$$
$$= (1 - \exp^{-\tau\gamma}). \tag{3.2}$$

From this, the probability mass function is defined over τ by taking the derivative as,

$$f(\tau) = \gamma \exp^{-\gamma\tau}. \tag{3.3}$$

From Eqs. (3.1) and (3.3), we have the average probability of a vertex infecting and information on the time of recovery, τ. Thus, total probability of transmission is

$$T = \int_0^\infty T(\tau) f(\tau) d\tau = \frac{\beta}{\beta + \gamma}. \tag{3.4}$$

From this, a percolation process is evaluated on through a degree distribution using probability generating functions.

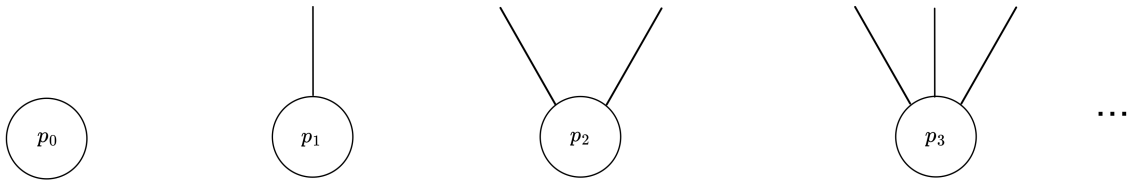

Figure 3.2: Contact degree distribution, $G_0(x)$, visual representation: For an arbitrary degree distribution, all probabilities of degree k, p_k, can be represented in visual vertex form. This stub list of vertices indicate the degree and associated probability.

3.3 Probability generating functions

Generating functions are power series that act similar to an array in a programming script: the constant position of a polynomial holds a value associated with the zero-th position. The constant term holds a value associated with the first position, the x term holds value associated with the second position, and this holds true for all other terms in the power series. Furthermore, when the coefficients of these generating functions are probabilities, this defines a probability generating function (PGF). At their foundation, PGFs provide mathematical structure to probability distributions associated with an arbitrary random counting variable, k. The probability associated with the count k is the coefficient for the x^k term.

For disease spread a contact network degree distribution defines the probability of a randomly chosen vertex having degree k or k neighbors, denoted as p_k [73, 99]. We assume the number of vertices in the contact network is infinite [73]. Therefore, we can define the PGF for a contact degree distribution as

$$G_0(x) = p_0 + p_1 x + + p_2 x^2 + ...,$$
$$G_0(x) = \sum_{k=0}^{\infty} p_k x^k. \tag{3.5}$$

Another way to describe G_0 is as the PGF that represents the probability of choosing a random vertex with degree k, as shown in Fig. 3.2. With this polynomial structure holding all the information on the probability distribution, there are three operations that provide specific information on the original degree distribution: derivatives, moments, and powers. First, applying k derivatives G_0, then evaluating at $x = 0$ defines

$$\frac{1}{k!}\frac{d^k G_0}{dx^k}\bigg|_{x=0} = p_k. \tag{3.6}$$

This extracts the k^{th} probability, p_k, from the PGF. Evaluating at $x = 0$ is essential to this process because for the k^{th} derivative makes the k^{th} value, $(k!)p_k$, the only constant left of the function.

When evaluating G_0' at $x = 1$ instead, the result is,

$$G_0'(1) = \sum_k k p_k = \langle k \rangle, \tag{3.7}$$

producing the first moment, or average degree, of the probability distribution. Taking the derivative multiplies all terms by k, similar to the computation for a weighted average.

Next up, we can derive the other moments of the probability distribution through the PGF framework. The n^{th} moment is then given by the n^{th} derivative of G_0

evaluated at $x = 1$,

$$\left[\left(x\frac{d}{dx}\right)^n G_0(x)\right]_{x=1} = \sum_k k^n p_k = \langle k^n \rangle. \tag{3.8}$$

These moments can be combined to compute centered moments like variances, skewness and other higher moments of the distribution.

Finally, raising a PGF to a power result in,

$$[G_0(x)]^2 = \left[\sum_k p_k x^k\right]^2 = \sum_{jk} p_j p_k x^{j+k}. \tag{3.9}$$

This establishes a PGF for the sum of degrees or neighbors for the vertices generated by G_0. Since G_0 is raised to the power of 2, Eq. (3.9) defines the PGF for the probability of picking 2 vertices with their degrees summing to $j + k$. In general, if m vertices are chosen from a network, the probability distribution of the sum of the degrees of those vertices is generated by $[G_0(x)]^m$.

Consequently, all these properties help derive an important distribution related to G_0, its *excess degree distribution*. The excess degree distribution defines the degrees of vertices reached when choosing a random edge to follow. To distinguish this from G_0, we denote the PGF for the excess degree distribution as G_1. Now, we know,

$$G_1(x) \propto \sum_k k p_k x^{k-1} \tag{3.10}$$

because the excess degree distribution is proportional to the degree k, multiplied by the original degree distribution. Intuitively, this means the higher the degree of a vertex, the more likely to chose an edge to their neighbor. With k as a multiplier, this resembles the derivative formulation, therefore, we assert

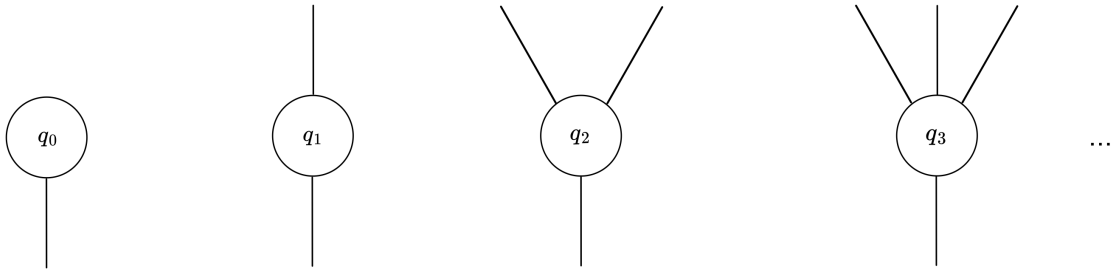

Figure 3.3: Excess degree distribution, $G_1(x)$, visual representation: For an excess degree distribution, all probabilities of degree k, q_k, can be represented in visual vertex form, where an imaginary vertex from G_0 is connected to the bottom of the randomly chosen edge. This stub list of vertices indicate the degree and associated excess degree probability.

$$G_1(x) = \frac{1}{\langle k \rangle} \frac{d}{dx} G_0(x) = \sum_k q_k x^k. \tag{3.11}$$

Figure 3.3 illustrates how the excess degree distribution provides the probability distribution for the number of first neighbors of an arbitrary vertex from the original degree distribution. To compute the number of second neighbors of the original degree distribution, we use both the original degree distribution and the excess degree distribution. Figure 3.4 intuitively depicts how to compose the number of second neighbors from the other two distributions. For every p_k in G_0 there is an associated excess degree distribution, defined by

$$\sum_k p_k [G_1(x)]^k = G_0(G_1(x)). \tag{3.12}$$

We note that when a PGF replaces the counting variable, x, for another PGF, we are composing PGFs. The composition defined in Eq. (3.4) defines how the sum

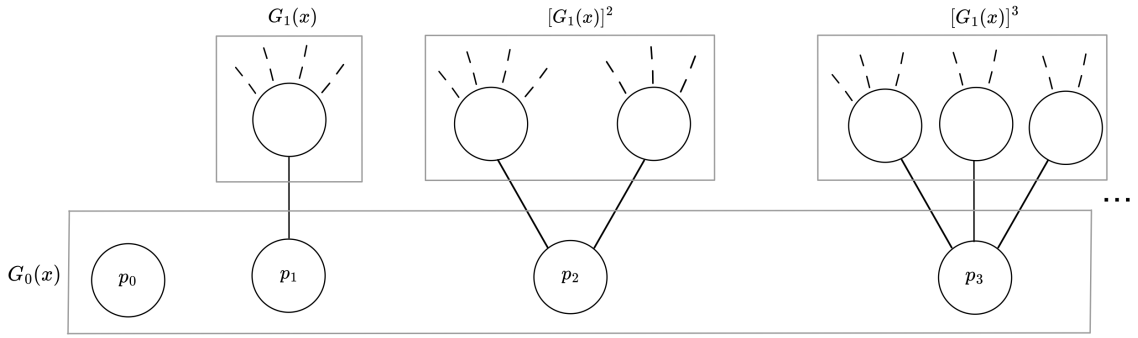

Figure 3.4: Second neighbor degree distribution visual representation: When evaluating the second neighbor degree distribution of a randomly chosen vertex, it is necessary to start at the original degree distribution. From there, the excess degree distribution builds on another layer of vertices as seen in this multi-layer stub list through a composition of functions.

of the excess degrees for each randomly chosen vertex produces the distribution for number of second neighbors.

Moreover, percolation can occur explicitly on the networks defined by $G_0(x)$ via the transmissability term, T, resulting in a composition of PGFs. Given an infectious vertex with degree k, we state the probability that the vertex infects ℓ neighbors as,

$$p_{\ell|k} = \binom{k}{l} T^l (1-T)^{k-l}. \tag{3.13}$$

This defines a binomial distribution of k independent trials where $p_{\ell|k} = 0$ if $\ell > k$.

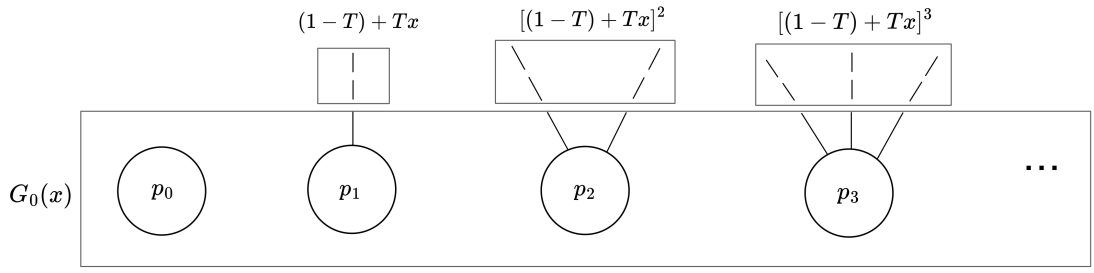

Figure 3.5: Transmission on $G_0(x)$ visual representation: To percolate an infection on the original degree distribution, $G_0(x)$ establishes the first layer of probabilities, then probability of transmitting, T, through certain edges is applied to each edge.

Integrating this probability into $G_0(x)$ establishes

$$\begin{aligned} G_0(x;T) &= \sum_{l=0}^{\infty}\sum_{k=l}^{\infty} p_k p_{l|k} x^l \\ &= \sum_{k=0}^{infty}\sum_{l=0}^{k} p_k \binom{k}{l} T^l (1-T)^{k-l} x^l \\ &= \sum_{k=0}^{\infty} p_k ((1-T)+Tx)^k, \end{aligned} \tag{3.14}$$

which defines the percolated $G_0(x)$ PGF. Consequently, $(1-T)+Tx$ represents a PGF with with probabilities $(1-T)$ and T from the previously defined binomial distribution. This establishes a composition of PGFs, which we illustrate with the visual representation of the percolated network PGF in Fig. 3.5. From the figure, we illustrate the binomial choice for each edge of the contact degree distribution. In the same manner that $G_0(x;T)$ is derived, we can determine $G_1(x;T)$. This PGF defines the probability distribution of the number of infections caused by a single node, otherwise known as the secondary case distribution [72].

This framework provides an efficient way to define a degree distribution in a

mathematical structure. Moreover, the derivations from the PGFs provides useful information about distribution descriptors, neighbors, and transmission spread. The subsequent sections show two ways in which degree distribution PGFs, in conjunction with a generalized counting PGFs, model different outcomes of disease spread.

3.3.1 Generational Spread Analysis

Generational spread analysis fills in a gap in the PGF framework, a time component. Noël *et al.* extend the PGF methodology in their 2009 work to model the sizes of epidemic generations through tracking temporal case counts, illustrated in Fig. 3.6 [74]. This figure shows the progressions of an infection from one vertex to the next. A vertex belongs to generation g if it became infected via a neighbor belonging to generation $g-1$. This property assumes an infinite-size random network drawn from a specific degree distribution, otherwise known as the configuration model [31]. Configuration models assume that each branch of subsequent infections are uncorrelated from each other. Therefore, each vertex in each generation can be treated as independent from all other vertices in its generation.

For each infected vertex in generation g, a piece-wise generating function describes the generation of cases that vertex will cause over the course of the epidemic [74]. This PGF is defined as

$$G_g(x;T) = \begin{cases} G_0(x;T) & (g=0) \\ G_1(x;T) & (g>0), \end{cases} \tag{3.15}$$

where $G_0(x;T)$ informs the degree distribution the of initial infected vertex, or 'pa-

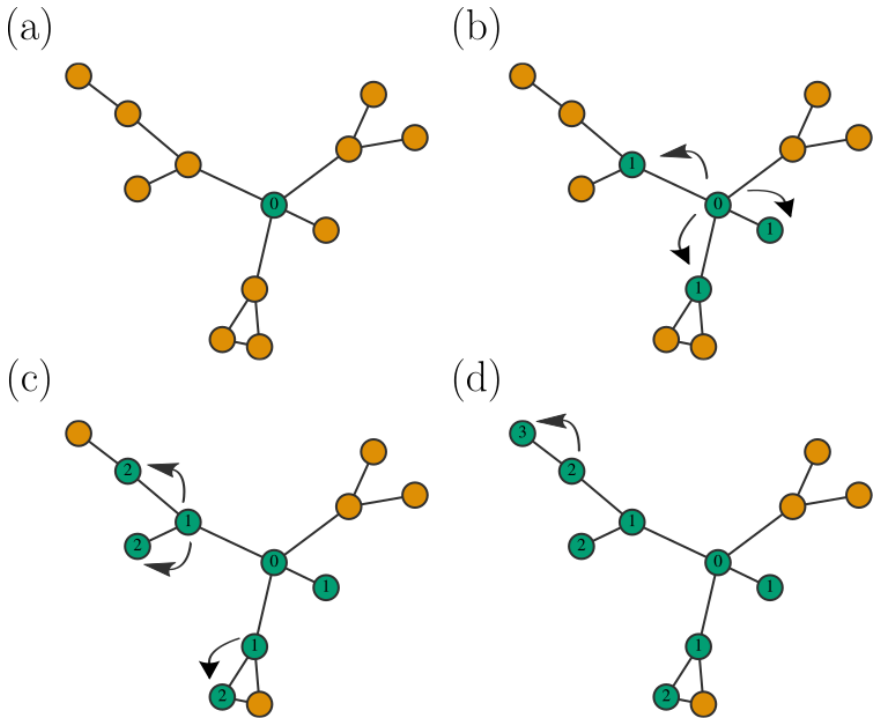

Figure 3.6: Generational infections: Over four generations, an infection spreads through a network. Each vertex's label corresponds to the epidemic generation in which it was infected. The initial infected vertex is in generation 0, any vertices they infect constitute generation 1, and so on. This figure is originally from work of Allen *et al.* [4].

.

tient zero'. $G_1(x;T)$ then defines the secondary cases caused by 'patient zero' for all subsequent generations.

Now, Noël *et al.* aim to understand the evolution of the cumulative case distributions. We define s as the number of cumulative cases at generation g, and m as the number of infectious vertices strictly belonging to g. Therefore, s is the sum of all m values for all generations, including generation g. The probability of s total infections by the end of g with m new infections during that generation is denoted as

ψ^g_{sm}. Furthermore, we define the associated PGF as

$$\Psi^g_0(x,y) = \sum_{s,m} \psi^g_{sm} x^s y^m \tag{3.16}$$

over all s and m.

In order to evaluate $\Psi^g_0(x,y)$, we must revisit the mechanisms of the branching processes on networks. Each new infectious vertex in generation $g-1$, totaling m', informs the distribution of new infections, which is generated by $G_{g-1}((1-T)+Tx)$. Consequently, this is equivalent to defining the probability of infecting m new vertices in generation g, given the pair (s',m') in generation $g-1$,

$$\sum_m P(m|s',m')x^m = [G_{g-1}(x;T)]^{m'}. \tag{3.17}$$

Stepping back for a moment, we know that being in state (s',m') in generation $g-1$ is given by the probability $\psi^{g-1}_{s',m'}$. The state (s',m') also produces m new infections in generation g, meaning the new state is $(s'+m,m)$. Thus, we redefine the distribution of s cumulative cases, and m new infections as

$$\begin{aligned}
\Psi^g_0(x,y) &= \sum_{s,m} \psi^g_{sm} x^s y^m = \sum_{s',m} \psi^g_{sm} x^{s'} (xy)^m \\
&= \sum_{s'm'} x^{s'} \sum_m \psi^{g-1}_{s'm'} P(m|s',m')(xy)^m \\
&= \sum_{s',m'} \psi^{g-1}_{s'm'} x^{s'} \sum_m P(m|s',m')(xy)^m
\end{aligned} \tag{3.18}$$

where $\psi_{s'm'}^{g-1} P(m|s', m')$ is the probability of m new infections occurring in state (s', m'). Finally, we use the equivalence in Eq. (3.17) to derive

$$\begin{aligned}\Psi_0^g(x, y) = &\sum_{s'm'} \psi_{s'm'}^{g-1} x^{s'} [G_{g-1}(xy; T)]^{m'} \\ = &\Psi_0^{g-1}(x, G_{g-1}(xy; T))\end{aligned} \tag{3.19}$$

defining a recurrence relation for $g \geq 1$ for the PGF being in the state (s, m) in generation g. We assume that $\psi_0^0 = xy$ when there is only one initial infectious individual, which leads to $\psi_{sm}^0 = \delta_{s1}\delta_{m1}$ [74].

From Eq. (3.19), we can extract the distribution of cumulative infections, s. This is achieved by taking the marginal distribution over y and evaluating $y = 1$ for all m of Eq. (3.16). Then, by summing over all m we achieve the PGF for the distribution of cumulative infections at each generation as,

$$\Psi_0^g(x, 1) = \sum_{s,m} \psi_{sm}^g x^s = \sum_s \sum_m \psi_{sm}^g x^s = \sum_s p_s^g x^s. \tag{3.20}$$

where p_s^g is the probability of having s cumulative case in g. Figure 3.7 showcases how accurate this framework is in comparison to the stochastic simulations framework explained in Sec. 1.2.2. An application of this framework will impose interventions to understand how mitigation tools defines various aspects of these distributions in Chapter 4.

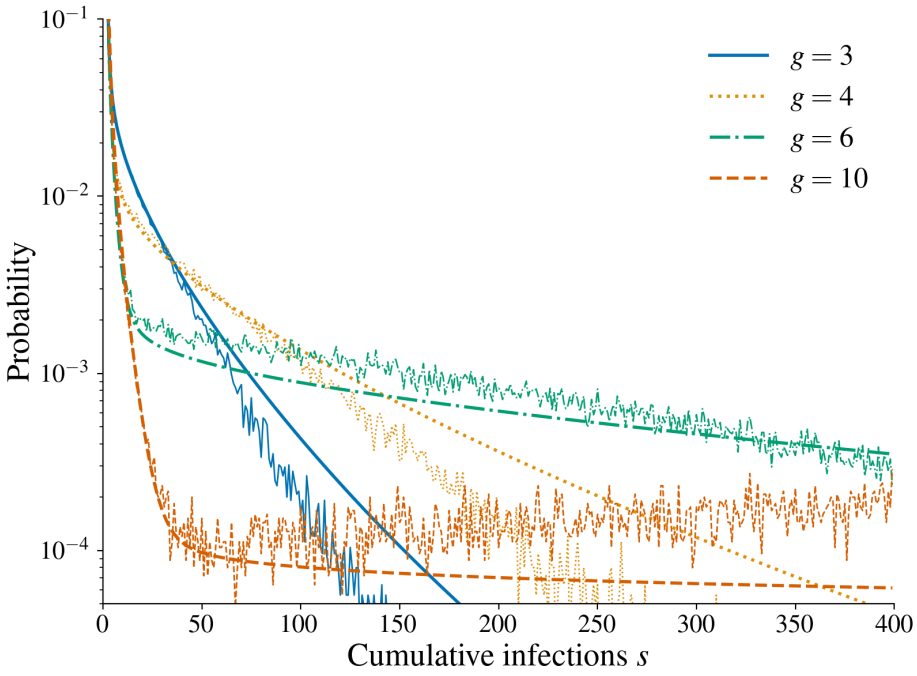

Figure 3.7: Time evolution of epidemics on a power-law network: The probability of having s cumulative cases by and during generation g for generations 3, 4, 6, and 10. This system modeling the cumulative case count distribution (smooth lines) for a modified power-law random network with a degree distribution defined by $p_k = k^{-2}e^{-k/10}$. The average degree of this network is $\langle k \rangle = 1.79$, along with its average excess degree of $\langle q \rangle = 3.04$. The distributions are validated by 75,000 simulations performed on 150 random network realizations with 10,000 vertices. The overall details of these simulations are mentioned in Chapter 4 and come from the work of Allen *et al.* This figure is originally from the work of Allen *et al.* as well [4].

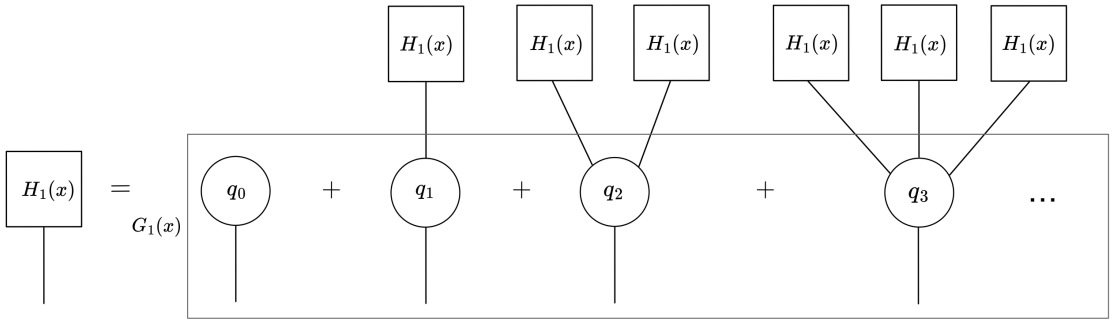

Figure 3.8: Finite component size distribution from a randomly chosen edge, $H_1(x)$ visual representation: From the composition of PGFs, we can determine the distribution of component sizes from a randomly chosen edge. The two PGFs that are composed together to established this distribution are $G_1(x)$ and $H_1(x)$. This provides a self-consistent relationship, since the excess degree distribution is connected to either another component or nothing.

3.3.2 Giant Component Analysis

Networks are defined as either connected or disconnected, meaning there is either path from on vertex to all others, or there is not [15]. Meaning a network can be composed of one component or many separate components. Component sizes can vary according to the governing degree distribution. Therefore, we can determine the distribution of finite component sizes and encode this into a PGF. Let the PGF for the distribution of finite component size when following a randomly chosen edge be $H_1(x)$. Figure 3.8 details how a component (square) from a randomly chosen edge is defined by the components connected to a vertex in the excess degree distribution. Figure 3.8 follows the same format as other composition examples, therefore, we define $H_1(x)$ as,

$$H_1(x) = xq_0 + xq_1 H_1(x) + xq_2[H_1(x)]^2 + ...$$
$$H_1(x) = xG_1(H_1(x)) \tag{3.21}$$

No matter what, an edge will lead to a vertex, which will either lead to more aspects of the component, or end at that vertex. Due to this fact, $H_1(x)$ can never have a constant term, meaning we cannot track components of size 0, hence why an x multiplier is added to the equation.

Similar to how we derive $H_1(x)$, we derive the PGF for the distribution of finite component size from a randomly chosen vertex, $H_0(x)$. Now, the PGF framework established infinite number of counting variables. While an infinite sized component could exist, it cannot be directly computed. We emphasize that $H_0(x)$ only defines finite component size and excludes a giant component, which is described as an infinite component. Figure 3.9 follows the same format at Fig. 3.8, however, we see $G_0(x)$ replaces $G_1(x)$. Hence, the equation defining $H_0(x)$ is

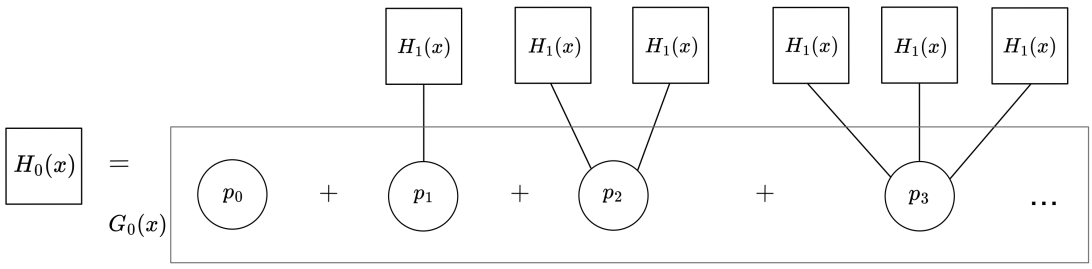

Figure 3.9: Finite component size distribution from a randomly chosen vertex, $H_0(x)$ visual representation: Similar to the visual for $H_1(x)$, we leverage the composition of PGFs to determine the distribution of finite component sizes from a randomly chosen vertex. $G_0(x)$ composed with $H_1(x)$ establishes the distribution of finite component sizes from a randomly chosen vertex, excluding the giant component.

$$H_0(x) = xp_0 + xp_1 H_1(x) + xp_2[H_1(x)]^2 + ...,$$
$$H_0(x) = xG_0(H_1(x)). \tag{3.22}$$

Remembering how $H_0(x)$ defines the distribution of component sizes exclude the giant component, the sum of the probabilities in $H_0(x)$ is,

$$H_0(1) = 1 - S. \tag{3.23}$$

This equation establishes the probability of randomly choosing a vertex in not the giant component. Therefore, S is the probability of randomly choosing a vertex in the giant component. We can alternatively state that S is the fraction of the network in the giant component because ware assuming an infinite network. Subsequently, this derivation can also be performed on $H_1(x)$, meaning

$$H_1(1) = u, \tag{3.24}$$

where u is the probability that a randomly chosen edge does not lead to the giant component. Furthermore, Eq. (3.21) evaluated at $x = 1$ produces,

$$H_1(1) = (1)G_1(H_1(1))$$
$$u = G_1(u), \tag{3.25}$$

which is another self-consistent equation defining the probability that a randomly

chosen edge does not lead to the giant component in the excess degree distribution. In a similar manner, evaluating $H_0(x)$ at $x = 1$ results in

$$\begin{aligned} H_0(1) &= (1)G_0(H_1(1)) \\ 1 - S &= G_1(u). \end{aligned} \tag{3.26}$$

Consequently, to satisfy this equation,

$$G_1(u) - u = 0, \tag{3.27}$$

means we are solving for the root of the polynomial, u. When assuming that $G_0(x)$ and $G_1(x)$ are equal to each other [34], u can then be directly used to solve for the proportion of the network in the giant component, S, from Eq. (3.26). This results in

$$\begin{aligned} 1 - S &= G_0(u) \\ 1 - S &= G_1(u) \\ S &= 1 - u. \end{aligned} \tag{3.28}$$

In a disease modeling context, this framework determines final outbreak sizes for an infinite population. For an example of the curves generated by Eq. (3.27), Fig. 3.10 provides three PGFs with negative binomial coefficients. The negative binomial is parameterized by R_0, average secondary cases and α, the dispersion parameter. Each curve has a unique parameter combination, as shown in the legend. We note that 1 is always a root because $G_1(1) = 1$, but if the curve crosses the x-axis prior,

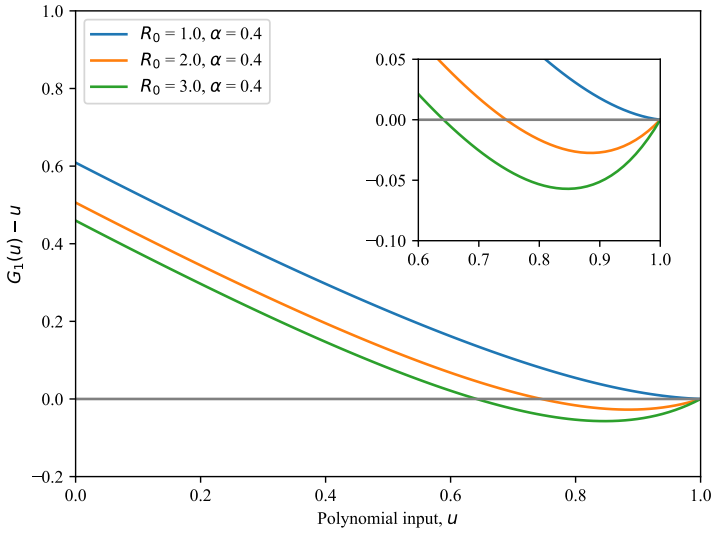

Figure 3.10: Solving $G_1(u) - u$ for polynomial roots: For a negative binomial distribution parameterized with the mean of the distribution and the dispersion parameter, we plot inputs of the polynomial versus the polynomial outputs. A polynomial root is defined as where the polynomial output is equal to 0. For three parameter combinations, we see the root shrink as the mean of the system increases. This in turn makes the giant component bigger.

the crossing determines another polynomial root. The sensitivity of this method is explored in Chapter 5.

Chapter 4

Temporal and probabilistic comparisons of epidemic interventions

Abstract

Forecasting disease spread is a critical tool to help public health officials design and plan public health interventions. However, the expected future state of an epidemic is not necessarily well defined as disease spread is inherently stochastic, contact patterns within a population are heterogeneous, and behaviors change. In this work, we use time-dependent probability generating functions (PGFs) to capture these characteristics by modeling a stochastic branching process of the spread of a disease over a network of contacts in which public health interventions are introduced over time. To achieve this, we define a general transmissibility equation to account for varying

transmission rates (e.g. masking), recovery rates (e.g. treatment), contact patterns (e.g. social distancing) and percentage of the population immunized (e.g. vaccination). The resulting framework allows for a temporal and probabilistic analysis of an intervention's impact on disease spread, which match continuous-time stochastic simulations that are much more computationally expensive. To aid policy making, we then define several metrics over which temporal and probabilistic intervention forecasts can be compared: Looking at the expected number of cases and the worst-case scenario over time, as well as the probability of reaching a critical level of cases and of not seeing any improvement following an intervention. Given that epidemics do not always follow their average expected trajectories and that the underlying dynamics can change over time, our work paves the way for more detailed short-term forecasts of disease spread and more informed comparison of intervention strategies.

4.1 Introduction

Monitoring the spread of COVID-19 is at the forefront of public health agendas as new variants emerge. Transmission across the globe has forced countries to mitigate the spread with their own combination of masking and social distancing [20], restrictions on mobility [1, 7], improved ventilation [92], contact tracing [51] and other local interventions. Even in neighboring regions, the diversity of interventions reflect differences in local policy, culture, differences in local forecasts, as well as different goals for interventions [98]. For example, some populations may attempt to minimize the expected number of COVID-19 transmissions while other may only wish to minimize the probability of overwhelming their healthcare system. Whether or not these

different objectives would lead to the same policies is unclear given the underlying randomness and uncertainty inherent to epidemic forecasting.

There are two important issues to consider when comparing forecasts of epidemic interventions: Forecasts should be *probabilistic* and *time-dependent* as disease spread is stochastic and heterogeneous [74, 4]. Temporal probabilistic forecasts must then be summarized by specifying given statistics, as well as a temporal window to target, chosen to capture the intended goal(s) of the intervention. And, since forecasts evolve, the relative effectiveness of two policies can itself vary over time. Altogether, comparing multiple intervention policies is not as simple as comparing the averaged effective growth rate of the epidemic.

Past work on intervention comparisons has studied how different policies such as lockdown strategies or physical distancing impact disease trajectory within a population [64, 97, 77, 24, 21]. Most of the comparisons in the literature, however, are based around the average of the stochastic (often simulated) outcomes or present confidences intervals for derived measures such number of hospitalizations or the effective reproductive number [64, 24]. In comparison, our philosophy is more similar to probabilistic forecasting in meteorology, where a cone of uncertainty of storm paths or expected rainfall are the targets. We argue that new summary statistics, which directly compare disease outcomes and their probability of occurring, need to be developed to account for the stochastic nature of disease trajectories.

In this chapter, we use a mathematical framework to track the distribution of cumulative and active cases in a networked population over the course of epidemic generations. When compared to simulations, these epidemic generations offer a surprisingly accurate proxy for the actual temporal dynamics of the epidemic [4]. We

extend this framework in Sec. 5.2 to allow temporal interventions that affect parameter values or contact structure from one epidemic generation to the next, thereby modifying the probabilistic epidemic forecasts over time. In Sec. 4.3, we present specific network interventions and offer a series of summary statistics chosen to capture the different possible goals of these interventions.

We demonstrate our approach to a specific case study in Sec. 4.4 where we compare targeted and random vaccination rollouts. Targeted vaccination is meant to immunize highly connected individuals (e.g. healthcare workers) that are at higher risk of receiving and passing the epidemic. However, this strategy comes at a cost and we assume that the targeted rollout of a vaccine must be slower than the random rollout of the same vaccine. Using our mathematical model and our summary statistics of temporal probabilistic forecasts we then ask: How fast must targeted vaccination be to outperform random vaccination? Do different metrics of intervention performance lead to the same answer? There are complex competition dynamics occurring between the epidemics unfolding on a contact network and interventions rolled out to affect this network (see Fig. 4.1). This work establishes a framework to study this dynamics and answer the previous questions. Section 4.5 outlines the generality of our approach, showcasing other types of interventions which can be modeled using our methodology.

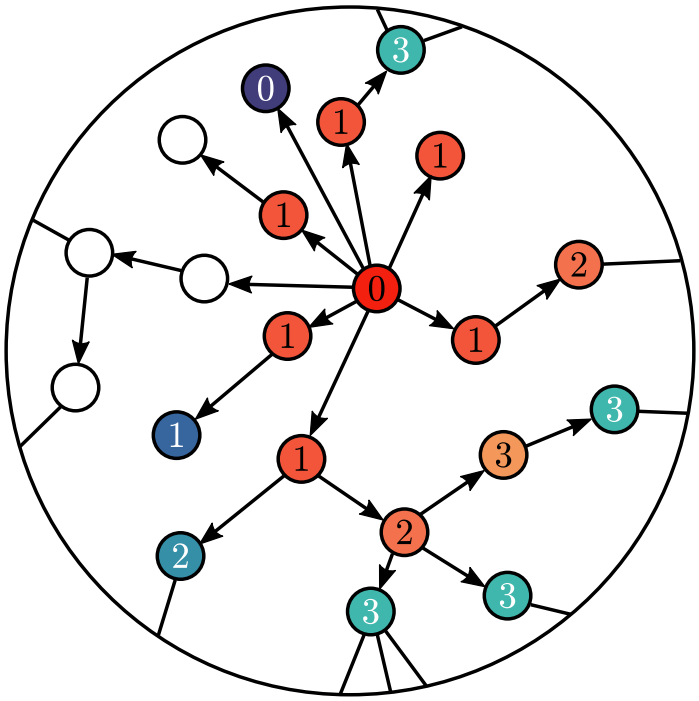

Figure 4.1: Schematic of generations of infection through a network with interventions. An initial node is infected during generation 0 (shown in deep red). Subsequent epidemic generations are represented in shades of red, with each node labeled in black by the generation in which it was infected. The blue shaded nodes were part of an intervention (e.g., vaccination), hindering the spread of the infection along that branch of the tree if the intervention preceded a potential transmission. Interventions are also temporal, shown in shades of blue and labeled in white by the epidemic generation when their intervention occurred. The branching dynamics of the resulting transmission tree are highly complex as the two dynamical processes compete, with the disease potentially spreading exponentially but slowing down as the intervention ramps up.

4.2 Theoretical Analysis

4.2.1 Assumptions

Our framework assumes that the spreading process of the disease being studied follows undirected percolation dynamics over a contact network and can therefore be

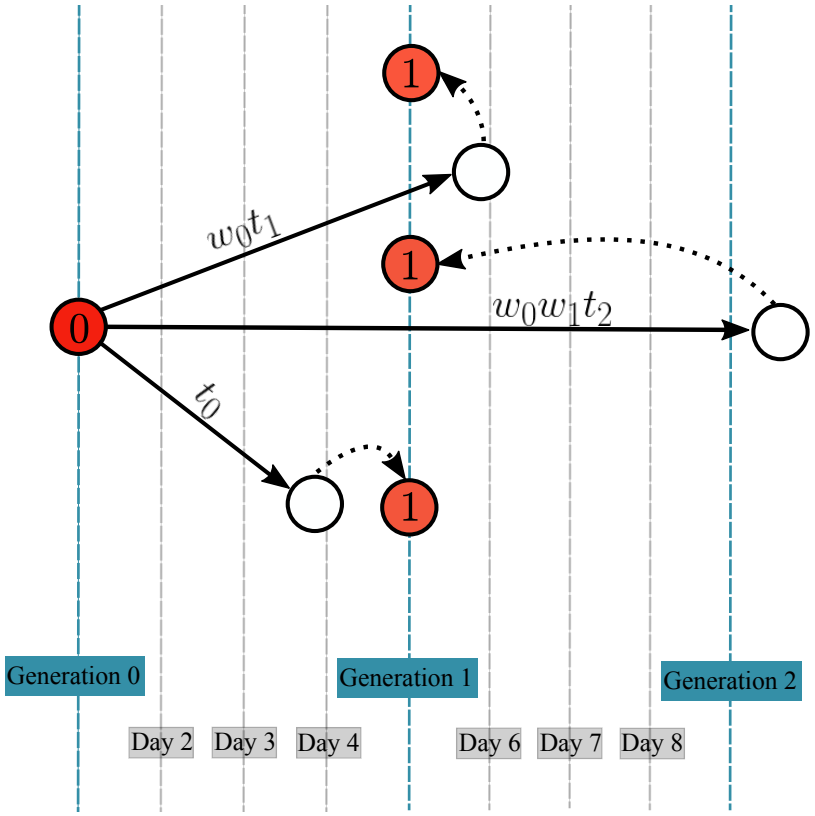

Figure 4.2: Mapping continuous-time dynamics to branching process generations. The process in which continuous-time disease spread is mapped to a discrete-time branching process is shown above. An infectious individual will infect a certain number of other individuals via a branching process, which is captured by the various transmission terms in Eq. (4.16). Once those individuals are identified, they are mapped to the next epidemic generation. For this specific example, we have an initial infectious individual (red node marked by generation 0), that infects three individuals at different probabilities of infection. If the transmission occurs in the same generational-time interval, here in the 0-th interval with probability t_0, the new case (bottom red node) becomes infectious at generation 1. When the transmission occurs during generation 1, the individual is conceptually mapped back to the start of generation 1 (top red node) and this occurs with probability w_0t_1. This probability is the probability of the 0-th generation passing multiplied by the probability of transmission occurring during the first generational interval. Likewise, there is a probability of two generations passing before a transmission occurs, with probability $w_0w_1t_2$, meaning the individual (middle red node) also mapped back to the start of generation 1. This mapping allows the analysis of continuous-time epidemic dynamics as a simpler discrete branching process.

analyzed as a branching process. Even though the underlying transmission dynamics occur in continuous time, we determine the probability of infection according to discretized generational time. This is represented in Fig. 4.2, where each solid-line arrow is a transmission event labeled with the probability of infection. We map these stochastic transmissions to a discretized epidemic generation. This discretization is shown in Fig. 4.2 with the vertical dotted lines representing time passing. At each generation, the branching process of transmissions from each infectious individual provides the new infectious individuals for the next generation. Even if transmission occurs in continuous time, the discrete-time mapping places individuals in the subsequent generation (sometimes underestimating time to transmission, sometimes overestimating it). The system is then updated and this process continues for the time-frame set. This assumption that transmission aligns with generational time makes analytical calculations and tracking of active cases easier even though it introduces small errors given that transmissions are pulled backward and forward in time, see Fig. 4.2 caption for an example case. This is an approximation of a spreading process [46] but was recently shown to provide accurate temporal forecasts when compared to continuous-time simulations [4].

In our case studies, we also assume that contacts in the population follows a geometric distribution. The aim of this assumption is to have a heterogeneous network of contacts. The geometric distribution is the discrete equivalent of the exponential distribution, which has been observed in real-world contact patterns [1-2]. With this distribution, we calculate the average number of secondary cases, R_0, to be 3. Though contact networks are inherently temporal, we here assume a static contact network except for the removal of connections due to network-based interventions.

When applying an intervention, specifically a vaccination strategy, we assume that vaccination offers perfect protection. Likewise, the intricacies of vaccine efficacy (e.g. waning of immunity or the need for multiple doses) will not be covered in this work but could be incorporated in the framework. Our goal is instead to provide a general model of disease spread and showcase how a few specific types of interventions can be included in temporal, probabilistic, and analytical forecasts. The software associated with our model is available at Refs. [2, 3].

Forecasts, in our framework, are defined as the time evolution of our branching process approach and will not be directly validated with data. While the final states predicted by our general approach have been previously validated with empirical data [1], data to produce temporal forecasts of interventions are not available. Further validation would require contact distributions, epidemiological parameters, and incidence rates within communities before and after interventions. Instead, we rely on simulations for validation.

4.2.2 Noël et al. probability generating function (PGFs) formalism

PGFs allow us to include inherent heterogeneity in epidemiological forecasting by calculating the probability distribution associated with specific network transmission trees. Generating functions offer elegant derivations of many statistical properties [99, 72].

For epidemiological forecasting purposes, the focus is on the PGF of the network

degree distribution, defined as

$$G_0(x) = \sum_{k=0}^{\infty} p_k x^k, \tag{4.1}$$

where the kth coefficient, p_k, is the probability of randomly choosing a node with degree k from the network. The average degree of the network, $\langle k \rangle$, is found by differentiating Eq. (4.1) and evaluating at $x = 1$,

$$G_0'(1) = \langle k \rangle = \sum_{k=0}^{\infty} k p_k. \tag{4.2}$$

This result is used to generate the distribution of potential transmissions, or the excess degree distribution,

$$G_1(x) = \frac{G_0'(x)}{G_0'(1)} = \frac{\sum_k (k+1) p_{k+1} x^k}{\langle k \rangle} = \sum_{k=0}^{\infty} q_k x^k. \tag{4.3}$$

The probability of reaching a node with degree k from a randomly chosen edge is represented by the coefficients $q_k \propto (k+1)p_{k+1}$ due to the fact that a node of degree $k+1$ is $k+1$ times more likely to be connected to a random edge than a node of degree 1. The node of degree $k+1$ then has k remaining edges to transmit through, which corresponds to the derivative and renormalization of the original PGF.

To incorporate the disease spread through the excess degree distribution, q_k, it is necessary to include, $p_{l|k}$, the probability of ℓ transmissions from a single infectious node, given that it has excess degree k,

$$p_{l|k} = \binom{k}{l} T^l (1-T)^{k-l}, \tag{4.4}$$

where T is the probability of transmission and is further explained in Sec. 4.2.3. Therefore, the number of infections caused by "patient zero" is equal to the probability of having degree k and transmitting the disease to ℓ of those k neighbors. This is defined as $G_0(x;T)$, given by

$$\begin{aligned} G_0(x;T) &= \sum_{l=0}^{\infty}\sum_{k=l}^{\infty} p_k p_{l|k} x^l \\ &= \sum_{k=0}^{\infty}\sum_{l=0}^{k} p_k \binom{k}{l} T^l (1-T)^{k-l} x^l \\ &= G_0\Big(Tx + (1-T)\Big). \end{aligned} \tag{4.5}$$

As $G_1(x)$ is derived from $G_0(x)$, so can $G_1(x;T)$ be derived from $G_0(x;T)$. In a static network, $G_1(x;T)$ represents the PGF for the probability distribution of the number of infections caused by a single node, i.e., the secondary case distribution.

Now, PGFs traditionally do not keep track of time as the branching process unfolds; however, Noël *et al.* developed a piece-wise generating function that tracks the branching process via generations [74]. Mathematically, for a static network, this is given by

$$G_g(x;T) = \begin{cases} G_0(x;T) & g = 0 \\ G_1(x;T) & g > 0, \end{cases} \tag{4.6}$$

where $G_0(x;T)$ defines the distribution for the first generation and $G_1(x;T)$ defines all future generations. In this work we expand on the framework laid out above to demonstrate the effect that temporal behaviors can have on the branching process.

Following Noël *et al.* [74], we calculate the cumulative case distribution. To do so,

we use a simple generation scheme illustrated in Fig. 4.1: Any transmission from a node infected in generation g is considered to be in epidemic generation $g+1$ regardless of the exact timing of the transmission event. From this, let s_g be the number of cumulative cases at generation g and let m_g be the number of infectious nodes strictly belonging to generation g. Note that in this way, $s_g = \sum_{g'=0}^{g} m_{g'}$. We denote ψ_{sm}^{g} the probability of having s total infections by the end of the g-th generation with m becoming infected (and thus being infectious) during that generation. We also denote

$$\Psi_0^g(x,y) = \sum_{s=1}^{\infty} \sum_{m=0}^{s} \psi_{sm}^{g} x^s y^m \tag{4.7}$$

the associated PGF. As demonstrated in Ref. [74], $\Psi_0^g(x,y)$ is derived via a recursive function for the probability of s_{g-1} total infections in generation $g-1$. Each new infection, m_{g-1}, in $g-1$ has its own possible transmission connections, $G_{g-1}(xy;T)$, incorporating all possible transmission events leading up to generation g. Mathematically, this is given by

$$\begin{aligned} \Psi_0^g(x,y) &= \sum_{s'=1}^{\infty} \sum_{m'=0}^{s'} \psi_{s'm'}^{g-1} x^{s'} [G_{g-1}(xy;T)]^{m'} \\ &= \Psi_0^{g-1}(x, G_{g-1}(xy;T)). \end{aligned} \tag{4.8}$$

Successive iterations of Eq. (4.8) from an initial condition (e.g., $\Psi_0^0(x,y) = xy$ for a single patient zero) then allows to compute ψ_{sm}^{g} at the desired generation g.

4.2.3 Formalism extension: altering transmission

Given a time point, intervention strategies can be implemented, altering the future dynamics of the disease spread. From Eq. (4.6), we generalize the piece-wise generating function to adhere to the intervention strategy being utilized. Given the type of intervention strategy, there can be multiple generations with an intervention implemented. So, to capture interacting temporal features of the disease spread and intervention, each epidemic generation is defined by its own PGF,

$$G_g(x; T_g), \tag{4.9}$$

as contact patterns change along with a new transmissibility expression, T_g, which will be derived in the following section. We represent this model in Fig. 4.1, where the branching process is dynamically slowed by an intervention rollout.

PGFs model a stochastic process which encapsulates the random nature of disease spread. The probability of a current infectious person causing a new infection, or the probability of transmission, is captured in T. We will follow Susceptible-Infectious-Recovered (SIR) dynamics which could depend on the time since infection t', the time-dependent transmission rate $\beta(t')$, and time-dependent recovery rate $\gamma(t')$. One could then calculate a general probability for transmitting before recovery, but the exact calculation is often model-dependent. We will follow most models and consider that transmission and recovery as simple Poisson processes occurring at fixed rate β and γ respectively. However, transmission occurs only if the contact is not immune,

which is true with probability $(1 - V_g)$, where V_g is proportion of the population that has been vaccinated by generation g. In other words, V_g is the cumulative proportion of the population vaccinated. Assuming infectiousness of a node in generation g lasts for some random time τ then the probability of the individual transmitting infection to another individual is

$$\begin{aligned} T(\tau) &= (1 - V_g)[1 - \lim_{\delta t \to 0}(1 - \beta \delta t)^{\tau/\delta t}] \\ &= (1 - V_g)(1 - \exp^{-\tau\beta}) \ . \end{aligned} \tag{4.10}$$

When evaluating the probability of a particular τ, the cumulative distribution function over τ is evaluated, shown by

$$\begin{aligned} F(\tau) &= (1 - \lim_{\delta t \to 0}(1 - \gamma \delta t)^{\tau/\delta t}) \\ &= (1 - \exp^{-\gamma\tau}) \end{aligned} \tag{4.11}$$

The above derivation uses the average rate of recovery, γ. Taking the derivative of Eq. (4.11) gives the probability mass function over τ,

$$f(\tau) = \gamma \exp^{-\gamma\tau} \ . \tag{4.12}$$

We can then compute the total probability of transmission by calculating the average probability of an individual transmitting before its recovery, given that the individual recovers at time τ. The average transmissibility for a generation, T_g, is therefore

$$T_g = (1 - V_g) \int_0^\infty T(\tau) f(\tau) d\tau = (1 - V_g) \frac{\beta}{\beta + \gamma}, \tag{4.13}$$

with this derivation following Refs. [44, 43]. The expression for T_g allows us to interpret the probability of transmission as the probability a transmission occurs first in a superposition of Poisson processes, transmission and recovery with rates β and γ respectively.

In our model, the passage of time to the next generation must also be included, which is determined by the product of the average excess degree, $q = G_1'(1)$, and the transmission rate yielding $q\beta$ [4]. Allen *et al.* confirms similarities in the mapping between continuous time and discretized generational time given this rate for passage of time. Section 3.3 discusses the continuous-time simulations that exhibit the validation of our generational approach. Treating this as another Poisson process, and allowing for interventions, we find the probability of a single person causing an infection leading to the next generation to be

$$t_g = (1 - V_g)\frac{\beta}{\beta + \gamma + (1 - V_g)q_g\beta}, \tag{4.14}$$

where again V_g is the cumulative proportion of the population vaccinated by at g, and where q_g is the generation-dependent average excess degree and is defined at Eq. (4.21). Similarly, the probability that the next generation occurs *before* a given person either transmits the disease or recovers is

$$w_g = \frac{(1 - V_g)q_g\beta}{\beta + \gamma + (1 - V_g)q_g\beta}. \tag{4.15}$$

To encapsulate the probability transmission given Eqs. (4.14) and (4.15) for each generation, we combine the probability of a single person causing an infection leading to the next generation with the sum of probabilities that the next generation occurs

before a particular transmission or recovery event,

$$\begin{aligned} T_g &= t_g + w_g t_{g+1} + w_g w_{g+1} t_{g+2} + ... \\ &= t_g + \sum_{\ell=g+1}^{\infty} \left(\prod_{\ell'=g}^{\ell-1} w_{\ell'} \right) \cdot t_\ell. \end{aligned} \tag{4.16}$$

This last expression closes our mapping of continuous-time SIR dynamics to a discrete-time branching process. The same recipe can be used to map other compartmental models to branching processes by including additional mechanisms. Alternatively, a SEIR with a fixed latent period of one epidemic generation can be implemented by setting $t_g = 0$ and $w_g = 1$ to delay transmission. Regardless, once transmissions dynamics are mapped to a discrete-time (or generational) branching process, our general framework is agnostic to the details of the transmission mechanisms.

4.3 Interventions

With an understanding of how the transmission process of this framework operates, we now aim to take individuals out of this process via intervention strategies. It is important to remember that we assume a node infected in generation g infects nodes that are mapped to be in generation $g + 1$. If an intervention strategy is implemented during a generation, it is assumed it would occur immediately at the start of that generation. Intervention strategies directly alter the numerical value of Eq. (4.16), then Eq. (4.8) is recalculated to update the probability of having

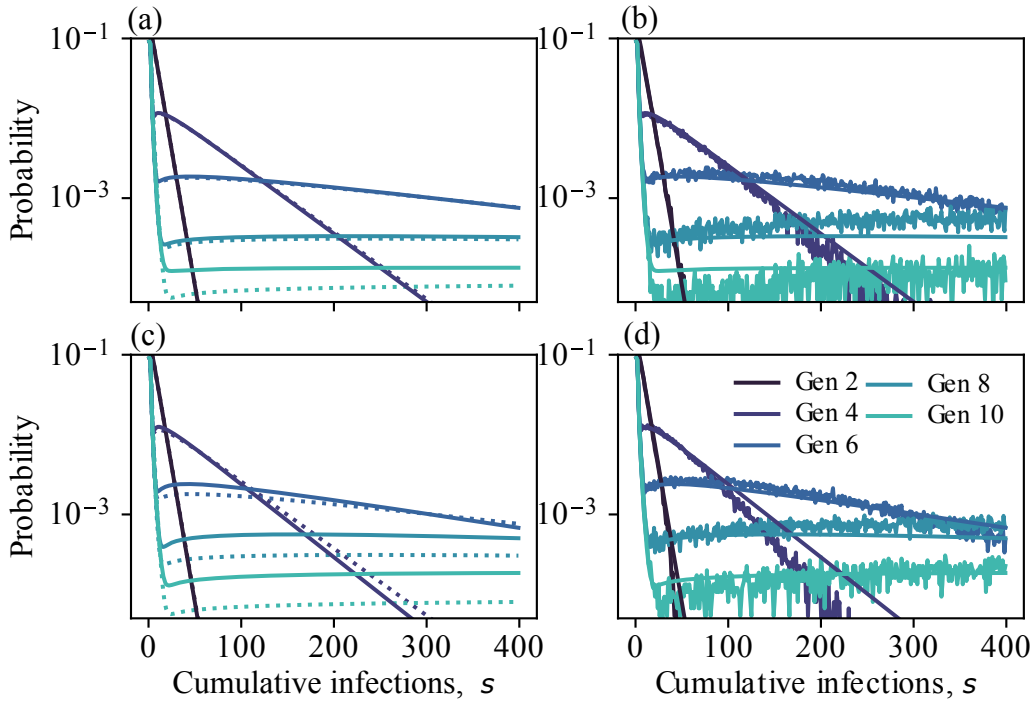

Figure 4.3: Random and targeted rollout comparison and validation. We use a geometric distribution defined by $p_k = 0.6^{k-1}(0.4)$, where $k = 1$, resulting in $R_0 = 3$. Each panel details probability distributions of cumulative infections at generations 2, 4, 6, 8 and 10. **Panel (a)** depicts the comparison between a non-intervened system (dotted lines) and a random rollout strategy of 0.5% of the population being randomly chosen to be vaccinated generations 4, 6, 8, and 10 (solid lines). By the end of generation 10, 2.0% of the population is vaccinated. **Panel (b)** depicts simulations of the random rollout vaccination strategy, which validates the modeled generations. **Panel (c)** depicts the comparison between a non-intervened system (dotted lines) and a targeted rollout strategy where the first 0.5% of highest degree individuals were chosen to be vaccinated at generations 4, 6, 8, and 10 (solid lines). **Panel (d)** depicts simulations of the targeted rollout vaccination strategy, which validates the modeled generations.

s cumulative infections and m active cases. The variables q_g and V_g that appear in Eq. (4.16), via t_g in Eq. (4.14) and w_g in Eq. (4.15), correspond with one of two types of intervention strategies. Respectively, the two types of strategies are: 1) Uniform or random interventions, where proportions of the population are not

susceptible to the disease. 2) Network interventions, which pertain to altering the degree distributions, such as targeted vaccination. When no network interventions are imposed on the system, q_g is kept consistent across all generations, after it is derived from the original excess degree distribution. Conversely, the value of V_g is kept at zero for all generations when there is a targeted intervention, since contacts around targeted vaccination are removed in q_g and no contacts can then lead to vaccinated nodes in this scenario. Here we focus on uniform interventions and network interventions, along with comparing them.

4.3.1 Uniform or random interventions

In this work, we consider the uniform intervention as a *random vaccination* strategy. This intervention strategy is implemented by randomly vaccinating susceptible nodes in the population with uniform probability [76]. The V_g term of Eq. (4.14) and Eq. (4.15) represents the probability of a node being vaccinated, along with the proportion of the population to be vaccinated at generation g. This quantity is therefore always a fraction between 0 and 1.

When a vaccination intervention is implemented at only one generation, meaning in a single intervention, the vaccinated population V_g changes as a simple step-function. Realistically, vaccination interventions are implemented over time and over multiple generations, which we can incorporate into our modeling framework by defining a *rollout* strategy. Under a vaccination rollout, the cumulative percentage of the population to be vaccinated is spread over multiple generations, slowly affecting the growth of the epidemic spread along each of its active generational branches. For multiple generations, we can state that the total proportion of the population vaccinated

over all generations, or cumulative percentage vaccinated, is given by

$$V_{total} = \sum_{g=1}^{\infty} V_g, \tag{4.17}$$

where each generational proportion vaccinated is defined as

$$V_g = \sum_{k=0}^{\infty} \delta_k^g p_k. \tag{4.18}$$

When implementing random vaccination, the entire network is uniformly vaccinated, resulting in

$$\delta_k^g = V_g, \tag{4.19}$$

for all k. It is important to note that random vaccination does not alter the general structure of the degree distribution or the average excess degree due to the condition set in Eq. (4.19).

Figure 4.3(A) showcases the difference in the probability distributions of cumulative infections on a system that has no intervention implemented (dotted lines), and one with a random rollout of 0.5% occurring at generations 4, 6, 8, and 10 (solid lines). Given that the intervention does not occur until generation 4, the distributions for generation 2 are exactly the same. For generations occurring after generation 4, the distributions begin to deviate from one another.

Figure 4.3(B) validates the extended PGF formalism for multiple interventions, with the theoretical distributions shown alongside numerical simulations following an event-driven, continuous time framework [4]. The analytical distributions show a bit of an overestimation at generation 10, which will be discussed in Sec. 4.3.3.

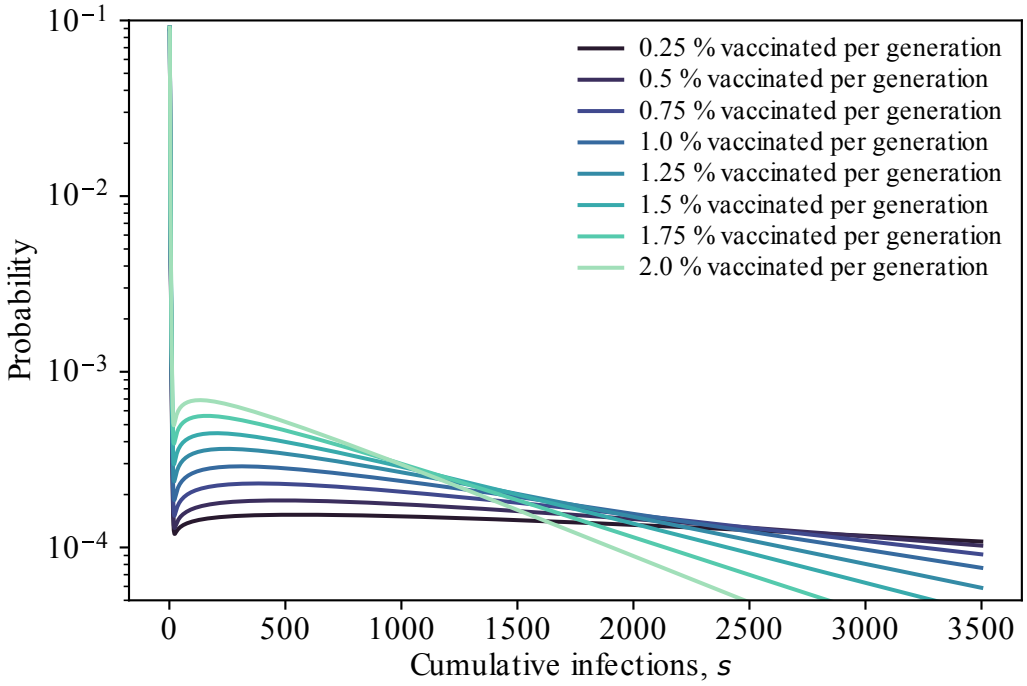

Figure 4.4: Flat distributions at generation 10. Given a geometric distribution defined by $p_k = 0.4^{k-1}(0.6)$, each line represents the probability distribution of cumulative infections at generation 10. The difference between the distributions is that the percentage of the population that were chosen to be vaccinated at generations 4, 6, 8, and 10 varies. The lower percentages per generation lead to flat distributions, whereas the higher percentages per generation provide distributions that have zero probability of cumulative infections past a certain point.

4.3.2 Targeted network interventions

To demonstrate a network intervention, in this work we show how a *targeted vaccination* strategy is implemented. The goal of targeted vaccination is to focus vaccination efforts on the group of nodes with the highest degrees in the network, or the individuals with the most contacts. This strategy results in reducing the impact of the individuals that have the most potential for creating a superspreading event, for an

example see [80].

Given a percentage of the population to vaccinate in g, as defined in Eq. (4.18), we start with degree classes $k' = k_{max}$ and vaccinate a fraction $\delta_{k'}$ of the degree class before moving to degree class $k' - 1$. To determine the fraction vaccinated for each degree class, we define

$$\delta_{k'}^{g} = \begin{cases} 1, & \text{if } \sum_{k=k'}^{\infty} p_k < V_g \\ \left(V_g - \sum_{k=k'+1}^{\infty} p_k\right) p_{k'}, & \text{if } \sum_{k=k'+1}^{\infty} p_k < V_g \\ & \text{and } \sum_{k=k'}^{\infty} p_k > V_g \\ 0, & \text{otherwise,} \end{cases}$$

where this non-uniform intervention will alter the degree distribution and average excess degree.

By the independence assumption of the configuration model, each neighbor will be vaccinated in generation g with probability equal to the probability that following a random edge leads to a vaccinated node, call this H_g. Thinking in terms of number of edges we thus compute

$$H_g = \frac{\sum_{k=0}^{\infty}(k+1)\delta_{k+1}p_{k+1}}{\sum_{k=0}^{\infty}(k+1)p_{k+1}}. \tag{4.20}$$

Therefore, the probability of a node being unvaccinated in g is equal to $1 - H_g$.

Now, the truncation of the degree distribution in g alters the average excess degree q_g, along with the coefficients of Eqs. (4.1) and (4.3). To determine the new q_g, we must recompute $G_g'(1)$. Multiplying this by the proportion of nodes that are

unvaccinated gives an updated q_g,

$$G'_g(1) = (1 - H_g)\left[\frac{\sum_{k=0}^{\infty} k(k+1)(1-\delta_{k+1})p_{k+1}}{\sum_{k=0}^{\infty}(k+1)(1-\delta_{k+1})p_{k+1}}\right] = q_g, \tag{4.21}$$

which is then used to derive the new T_g for a given g.

Similar to random vaccination described in Sec. 4.3.1, targeted vaccination can be implemented via one instance of vaccination, or multiple. Figure 4.3(C) shows the difference between a non-intervention strategy and a targeted rollout vaccination scheme of 0.5% at generations 4, 6, 8, and 10. A rollout strategy is conducted in the same manner for both random and targeted vaccination. Similar to random vaccination, the non-intervention leads a lower probability of seeing 100-400 cases than targeted (or random) vaccination. Does that mean the non-intervention is better? This question is answered in Fig. 4.4, which showcases how the weaker the intervention, the flatter the cumulative case distribution. These flatter distributions allow for there to be some chance of infecting more individuals over time. Figure 4.4 also shows that the stronger intervention, the more probability mass accumulates towards the smaller cumulative infection counts. This explains why interventions appear to do worse than no intervention at smaller values of cumulative case count. Even though this figure utilizes a targeted rollout vaccination strategy the same relationship holds for random vaccination strategies.

4.3.3 Validation via simulations

The simulations shown in Figs. 4.3(B) and (D) used to validate the theoretical framework were performed using an event-driven, continuous time approach on 150 distinct networks of 20,000 nodes with 500 simulations run on each network. This totals to 75,000 simulations per validation.

The analytical distributions of infection under different vaccination strategies capture the relationships between generations of infection, but tend to overestimate the number of cumulative infections compared to the continuous time simulations. This is due to a few factors; primarily, the finite-size effects of simulations on networks with 20,000 nodes, which support faster computational time but results in a sharper decrease in the size of epidemic generations than are captured by the branching process model, as discussed in Ref. [4]. Nonetheless, the important behavior of the distributions are captured in relation to one another, and across different vaccination strategies, allowing for comparison and ranking of the effects between strategies regardless of slight numerical precision errors.

Another source of discrepancy between the model and continuous time simulations is the inability of the model to account precisely for already-infected nodes by the time of an intervention. This quantity is estimated in Eq. (4.16), but may result in slight differences to the continuous-time simulations under which nodes who are already infected or recovered and are identified for targeted vaccination are not excluded, and are just ignored. This problem arises more for targeted vaccination efforts than random, since nodes in the targeted high degree classes are the same nodes that are likely to have been infected early in the spreading process. The results observed

in simulation may experience a reduced disease burden on the population than the theoretical model which assumes an infinite supply of these high-degree nodes, because the simulation on a finite network has already burned through its supply of high-degree nodes, rendering them recovered before the vaccination intervention.

4.3.4 Comparison of interventions

A simple comparison for differing interventions is a direct comparison of their cumulative probability distributions at a specific generation. Beyond this, there are metrics derived from the cumulative probability distributions that provide valuable information for decision makers.

Average cases First, we look at the expected number of cases over time, which we can denote as $\mathcal{X}^g_{mean}$. This corresponds to the typical approach using deterministic models that track the expected state of epidemics. Mathematically, this is defined by

$$\mathcal{X}^g_{mean} = \sum_{s=1}^{\infty} \sum_{m=0}^{s} s\psi^g_{sm} = \sum_{s=1}^{\infty} s\psi^g_s, \tag{4.22}$$

where ψ^g_s is the probability of s cumulative infections up to generation g.

Best - worst case The second metric looks at the worst case scenario over time as a measure of the underlying heterogeneity of possible epidemic sizes. This allows us to quantify what is the largest epidemic that has a realistic probability of occurring (set by some threshold probability, p_t) and therefore to select policies that offer the "best worst-case scenario" for robust decision making [53]. We derived this metric, denoted by $\mathcal{X}^g_{worst}$, by determining the corresponding cumulative case value, s', that

has probability p_t. Mathematically this means,

$$\mathcal{X}^g_{worst} = \underset{s'}{\operatorname{argmin}} \left\{ \left| \sum_{s=s'}^{\infty} \psi^g_s - p_t \right| \right\}. \tag{4.23}$$

Critical level of cases For the third metric, we look at the probability of being at or above a critical level of cases, c, defined by $\mathcal{P}^g_{flat}$. This metric is meant to capture the common goal of *flattening the curve* to avoid overwhelming the healthcare system [14]. From the cumulative probability distribution for cases, we derive

$$\mathcal{P}^g_{flat} = \sum_{s=c}^{\infty} \psi^g_s. \tag{4.24}$$

Minimal effect Finally, we have a metric that measures the probability that a realization of an epidemic with an intervention is actually worse than a realization of the same epidemic in the same population without the intervention, $\mathcal{P}^g_{worse}$. This summary statistic is meant to capture the fact that some interventions might have minimal effect on the expected spread of the disease which can easily be overshadowed by the intrinsic randomness of epidemics. Mathematically we define this as

$$\mathcal{P}^g_{worse} = \sum_{s=1}^{\infty} \bar{\psi}^g_s \chi^g_s, \tag{4.25}$$

where $\bar{\psi}^g_s$ is the probability of s cumulative cases when an intervention is implemented, and where $\chi^g_s = \sum_{i=1}^{s} \psi^g_s$ the probability of having less than or equal to s cumulative cases when there is no intervention implemented.

All of the metrics above provide varying emphasis on the information from the probability distributions for cumulative cases.

4.4 Case study: Random vs targeted vaccination

Given multiple vaccination strategies, choosing the best strategy involves comparing the impacts on the epidemic spreading process given some comparison criteria.

One could compare the random rollout strategy against targeted rollout strategy in Figs. 4.3(B) and (D). However, given the infinite distributions computed, we cannot see much of a difference in later generations between the two strategies unless we display more than 400 cumulative infections. This is also due to the fact that is showcased in Fig. 4.4, where the distributions could cross over each other past 400 cumulative infections.

When deciding on the best course of action and only considering the probability distributions for cumulative case counts, our results depict similar outcomes for a random and targeted vaccination strategy. With the same percentage of the population vaccinated, targeted rollout scheme proves slightly more effective, when focusing on larger cumulative cases in generations 2, 4 and 6. If the goal is to stop the spread of a disease early on, say by generation 6, a targeted rollout with a high percentage (V_g) per generation utilized at the given intervention generations needs to be implemented according to this model. This is only an example of a strategy determined by the distributions of cumulative cases. Other calculations can be performed on probability distributions of cumulative cases, which can inform decision makers on different courses of action, depending on desired goals in the beginning stages of an epidemic.

In the previous paragraphs, a comparison of distributions provides an overall com-

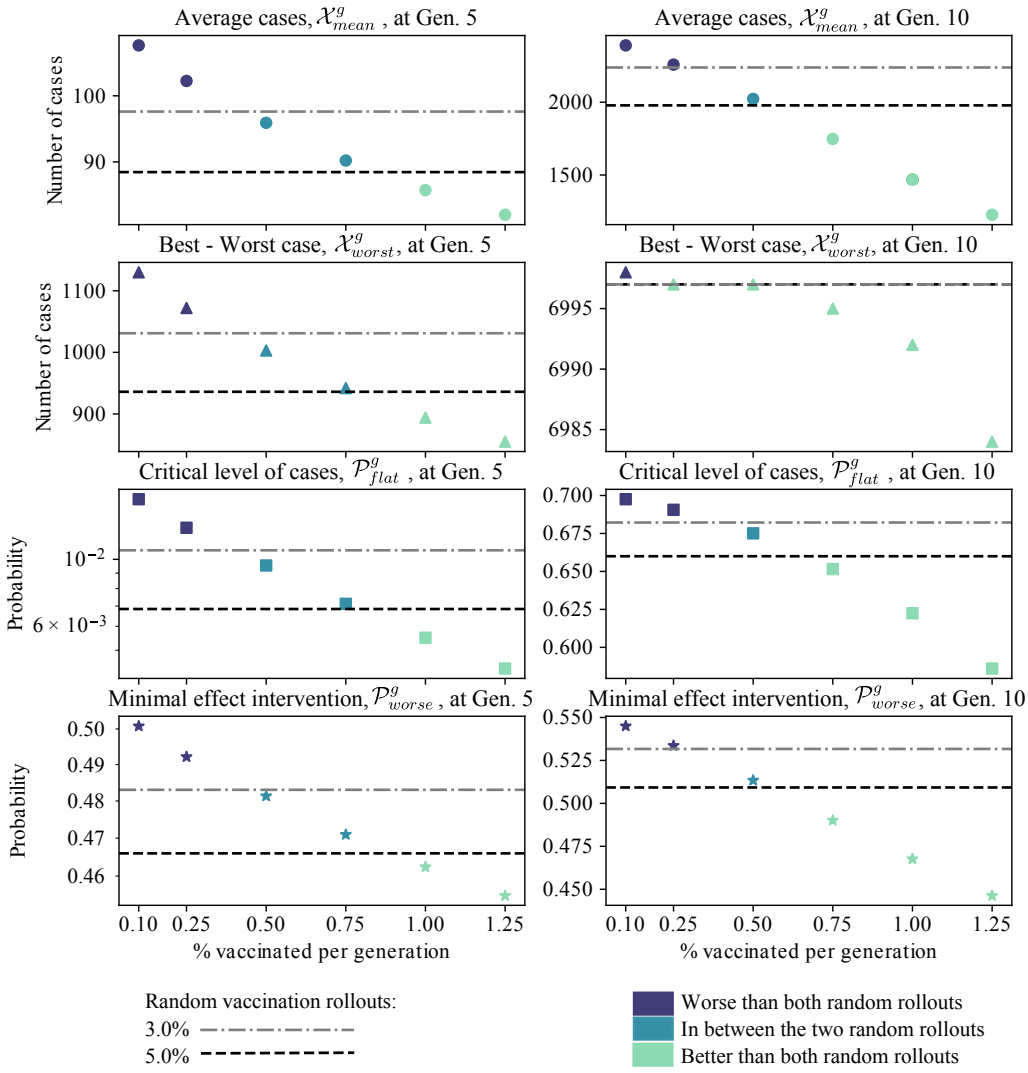

Figure 4.5: Varying targeted vaccination metrics compared to two random vaccination metrics. Given 0.10, 0.25, 0.50, 0.75, 1.00, and 1.25% targeted rollouts per generation (rollout occurring at generations 4, 6, 8, and 10) cumulative case probability distributions, the metrics defined in Sec. 4.3 are computed and along with the metric for a random rollout at 3.0% (dash-dot line) and 5.0% (dash-dash line). The differing colors of the makers represent whether the given targeted rollout is worse than, in between, or better than the two random rollouts, as shown in the legend. The threshold for the best-worst case metric is 10^{-4}. The critical level of cases is defined as 500 cases.

parison, however, we are able to calculate other metrics of comparison for differing vaccination strategies. The metrics defined in Sec. 4.3 appear in Fig. 4.5, which displays all of the metrics for targeted rollout with six different vaccination percentages ranging between 0.1 and 1.25%. Thus there are six different cumulative vaccination strategies represented in the figure. These percentages are applied to generations 4, 6, 8, and 10, hence total proportion vaccinated is cumulative; *e.g.* in the 0.1% case, 0.4% of the population will be vaccinated at generation 10. The horizontal lines represent random rollout vaccinations at 3.0% and 5.0%, which were rolled out at the same generations as the targeted strategy.

Figure 4.5 depicts a decrease in each metrics quantities as the percentage increases for a targeted rollout. For each metric, we observe that targeted vaccination with a 0.1% at each rollout generation is the only percentage that provides higher cases and probabilities for all the metrics of both random rollout strategies. All of the metrics are on varying scales, yet each one follows the decreasing trend as stronger targeted rollouts are applied. The scales for between generation 5 and 10 are drastically different as well, showing how a targeted vaccination increase of 0.25% drops the given summary statistic much more in generation 10 than generation 5. The 0.1% and 0.25% targeted rollout metrics are larger than the 5.0% random rollouts for almost all the measures at generation 10, however, the difference in resources between vaccinating a total of 0.4% or 1.0% is much smaller than vaccinating a total of 15% or 25% of the population to have a similar effect. The 0.5% and 0.75% targeted rollouts sit in the region between the two random rollouts for all the measures at generation 5. At generation 10, all of the metrics have the top three strongest targeted rollouts below both random vaccination strategies, along with the random vaccination lines moving

closer together on their given scales. We can even see the random vaccination lines being virtually the same at generation 10 given the scale of the number of cases. These calculations allow for decision makers to evaluate what strategy will best achieve their prioritized goal during an outbreak.

Overall, we find that there are three important factors influencing how fast a targeted roll-out must be to outperform a faster random rollout. First, the answer obviously depends on the speed of the random rollout itself. This intuitively makes sense due to the nature of targeted vaccination focusing on the individuals with the most connections. Second, it also depends on the desired temporal window: A targeted rollout of 0.75% per generation performed worse than a 5% random rollout per generation in all metrics at generation 5, but outperformed in all metrics by generation 10. Third, the preferred metrics of performance also influences the evaluation of interventions: Compared to a random rollout of 5% per generation, a targeted rollout of 0.75% per generation minimizes all the metrics. However, one should always consider relative differences between random and targeted rollouts to fully inform decision making.

4.5 Discussion

Once a decision has been made on whether or not to implement an intervention, the question of which strategy to use arises. Without a comparison of intervention strategies, decision makers may be lead to choose a scheme that does not mitigate the greatest concerns of their communities. Given the analytical derivations from this work, it is apparent that a targeted vaccination strategy has a faster impact on the

spread of disease than a random vaccination strategy. The choice between a single instance intervention and a rollout of interventions depends on the resources of the community under consideration. Some of the vaccination strategies that only intervene in one generation may not be feasible because there is simply not enough time to vaccinate 0.15% of the population during a generational window, let alone 3.0% for multiple windows. This fact, along with the difficulty of determining the super-spreading events in a community, makes targeted vaccination harder to coordinate than random. The relative costs of different strategies and which strategy makes the most sense in terms of resources and time is not within the scope of this chapter, but are important aspects for public health officials to consider.

Although this work does not evaluate all the intricate parts of implementing various intervention strategies, it successfully captures the stochastic nature of disease spread and the heterogeneity of contact patterns and human behaviors. Due to its generational time aspect, this temporal and stochastic model removes some of the assumptions in other forecasting models, which aim to derive random disease spread, along with the impediments to the spread, over time. Another advantage to this analytical model is the transmission expression defined in Sec. 4.2.3 has the flexibility to accommodate intervention strategies other than uniform or network interventions. Equations (4.14) and (4.15) can accommodate interventions such as treatment and transmission based interventions. Values for γ could depend on therapeutics, or a number of other types of treatments while values for β could depend on masking, ventilation improvements, social distancing, or testing. Altogether, Eq. (4.16) provides a flexible approximation to account for multiple interventions, and even combinations of interventions, in probabilistic forecasts. Comparing interventions is a multidimen-

sional problems, and therefore so is the design of interventions. Future work should include testing other intervention strategies, along with combining multiple strategies as we have seen happen around the world. Public health tools and forecasts need to be as heterogeneous and complex as the epidemics they aim to control.

Chapter 5

Sensitivity analysis of stochastic polynomial roots, and its application to epidemic forecasting and random graphs

Abstract

Probability generating functions (PGFs) help extrapolate useful information from the network degree distributions they generate. From an arbitrary degree distribution, PGFS facilitate derivations of the excess degree distribution, distribution of finite component sizes and giant component size. A giant component, or infinite size component that takes up a proportion of a network, are useful in understanding the propagation of a spreading process through a network. This work performs a sensitivity

analysis with condition numbers of polynomial roots from PGFs to understand giant component variation. We analyze two distributions generated by PGFs: a negative binomial distribution, and an Erdös-Rényi random graph. Our intuition is confirmed that the most sensitive regimes are those that produce the smallest possible giant component. From this framework, we discuss the implications for other PGF and branching process sensitivity analyses.

5.1 Introduction

Sensitivity analyses enable scientists to evaluate the bounds and scope of the models they use. As new methodologies come to light, their parameter spaces must be explored to understand any peculiarities from perturbed inputs. Global sensitivity analyses (GSAs) are available to assess models and their parameter spaces for this exact purpose [84, 100]. From scatter plots to latin hypercube sampling, GSAs can address the sensitivities for a range of model types. The type of model used in this chapter is a probability generating function (PGF). PGFs are stochastic models, which encode probability distribution values as coefficients of a power series. This power series or polynomial, carries information for the probability distribution, such as averages and other distribution descriptors [99]. Condition numbers, a variation of the elementary effects or the Morris method sensitivity analysis, commonly assess the sensitivity of polynomials [68, 52]. This chapter focuses on a condition estimator for polynomial roots of PGFs, which help derive the size of a giant component.

First, to understand the inputs of the model, we define a generalized network in terms of a probability distribution. As mentioned previously, a network defines

a collection of vertices and edges. Two vertices are connected if there is an edge between them. A *vertex*, or node, is said to have degree k when it has k neighbors or is connected to k other vertices by edges. Moreover, a degree distribution defines a probability distribution for the degree of a randomly chosen vertex. Consequently, the degree distribution values are the inputs of the PGF, otherwise known as its coefficients. These coefficients are subject to perturbations in the sensitivity analysis.

While there are multiple types of outputs for this model, the work focuses on the size of the giant component. Remember, a network is defined as either connected or disconnected, meaning all vertices are connected to all others or not. When a network is disconnected, it is made up of at least two connected components. Deriving the distribution of component sizes with the help of the original degree distribution, we describe if a randomly chosen vertex is in a finite component, or the giant component, which is infinite in size. A particular root value of the polynomial defined by the PGF, discussed in detail in Sec. 5.2.2, relates to the size of the giant component. Therefore, the variations in input values will affect this root value and giant component size.

Various spreading processes can occur on networks, for example, disease spread. The giant component inform scientists on the largest proportion of a population that could be infected by a disease. The sensitivity of the giant component is unknown and essential to evaluate, especially for epidemiological purposes. This chapter aims to evaluate the sensitivity of two distributions. The first being a negative binomial distribution, which informs epidemiological forecasting through final outbreak sizes. The second case evaluates an Erdös-Rényi random graph, which establishes the effect of varying levels of percolation on random graphs. Being the first sensitivity analysis of a PGF model, this sets the stage for sensitivity analyses for other PGF outcomes

used in epidemiological and non-epidemiological applications.

5.2 Methods

5.2.1 Assumptions

The assumptions around PGFs include that the degree distribution encoded in a PGF assumes a network of infinite size. This framework is not temporal. To determine the giant component or final outbreak size of a network, we define the network encoded in $G_0(x)$, and assume $G_0(x) = G_1(x)$. This assumption simply makes the polynomial root and outbreak size sum to 1. For the negative binomial case study in this work, we assume there is not a binomial choice with a transmissibility term, T, performed on an underlying contact network [34]. We also assume that for the negative binomial case, the largest k value is 20. However, for the Erdös-Rényi random graph case, we assume the largest k value is 10, given the small support of the Poisson distribution.

5.2.2 Probability generating functions

As detailed in Sec 3.3, PGFs are used to encode probability distributions in a compact way: as the coefficients of a formal power series [99]. This encoding allows for scientists to compute many properties of a distribution [99]. For an arbitrary network or graph applied to a PGF, there is a relevant random variable which indicates the probability that a vertex has degree k, we denote this p_k. This degree distribution is generated by

$$G_0(x) = \sum_{k=0}^{\infty} p_k x^k. \tag{5.1}$$

The number of connections of the first neighbor to a randomly chosen vertex is not governed by $G_0(x)$. Thus, the neighboring vertex itself is not randomly chosen by $G_0(x)$. To randomly chosen this neighboring vertex, we define the excess degree distribution, which describes the probability of following a randomly chosen edge to a vertex with degree k. We define a different PGF for this distribution, where individuals with k connections are k times more likely to be chosen by this process. Therefore, we generate the excess degree distribution with [73]

$$G_1(x) = \frac{\sum_k (k+1)p_{k+1}x^k}{\sum_k (k+1)p_{k+1}} = \frac{1}{\langle k \rangle} G_0'(x) = \sum_{k=0}^{\infty} q_k x^k, \tag{5.2}$$

which is normalized by the average degree, given by

$$\langle k \rangle = \sum_{k=0}^{\infty} k p_k = G_0'(1). \tag{5.3}$$

Consequently, we have detailed the PGFs that generate the degree distributions of a randomly chosen vertex, $G_0(x)$ and a randomly chosen edge or the excess degree distribution, $G_1(x)$. Moreover, the analysis for this chapter will focus on the giant component of a graph, which is derived via the distribution of the size of components arrived at by following a randomly selected edge, $H_1(x)$. Remember, this distribution excludes the possibility of an edge leading to the giant component. Nevertheless, we can directly solve for the giant component from the information encoded in $H_1(x)$ [73]. The distribution of component sizes when following a randomly chosen edge is generated by

$$H_1(x) = xG_1(H_1(x)). \tag{5.4}$$

This self-consistent equation is derived by noticing that $H_1(x)$ is equal to the proba-

bility of a following a randomly chosen edge, $G_1(x)$, to a component of a given size, which again is defined by $H_1(x)$. Similarly, the PGF describing the distribution of component sizes for a randomly chosen vertex is given by

$$H_0(x) = xG_0(H_1(x)). \tag{5.5}$$

All the distributions laid out above do not include the giant component, or the component of infinite size. Due to this fact, it is known that $H_1(1) \neq 1$, compared to $G_0(1)$ and $G_1(1)$ which do sum to 1. Furthermore, this leads to $H_1(1) = u$, where u is the probability of not following an edge to the giant component [73]. Therefore, Eq. (5.4) can be rewritten as

$$u = G_1(u), \tag{5.6}$$

when evaluating for $x = 1$ [73]. In a similar manner, when Eq. (5.5) is evaluated at $x = 1$, we derive

$$\begin{aligned} 1 - S &= G_0(u), \\ S &= 1 - G_0(u). \end{aligned} \tag{5.7}$$

For computational simplicity, it is assumed that $G_0(x) = G_1(x)$, meaning the degree distribution for the initial individual is equal to the degree distribution of all other individuals in the network [34]. Therefore, to calculate the proportion of the graph

the giant component occupies, S, we solve

$$S = 1 - G_1(u) = 1 - u. \tag{5.8}$$

The polynomial root, u, and giant component proportion, S, is the outcome for this sensitivity analysis.

5.2.3 Statistical condition of polynomial roots

The core idea of solving PGFs for giant component analysis is to determine the roots of Eq. (5.6). When dealing with noisy data, solving Eq. (5.6) is subject to the statistical condition of the solution. This is a measure of how sensitive the true solution is when perturbations are applied to the coefficients. Laub and Xia establish a framework, a statistical condition estimation (SCE), to assess exactly that [52].

The general steps for the variation of the Laub-Xia algorithm for this work goes as such: (i) we solve for the real roots, $\mathbf{x}$, of the polynomial $p(x)$ between $[0, 1]$. (ii) In order to assess sensitivity, we generate a matrix of random values to be added to each coefficient for z trials. (iii) Following the samples of z trials, we compute the perturbed roots, $\tilde{\mathbf{x}}$, for the perturbed polynomial, $\tilde{p}(x)$. (iv) Finally, the condition number for the root between $[0, 1]$ is given by the ℓ_2 norm of the component-wise division.

The Laub-Xia algorithm leverages the results from linear algebra, one of which being how the eigenvalues of the companion matrix of polynomial coefficients corre-

spond to the roots. The companion matrix of a polynomial is defined as

$$\begin{pmatrix} 0 & 0 & 0 & ... & p_0/p_k \\ 1 & 0 & 0 & ... & p_1/p_k \\ 0 & 1 & 0 & ... & p_2/p_k \\ 0 & 0 & 1 & ... & p_3/p_k \\ ... & ... & ... & ... & ... \end{pmatrix}, \tag{5.9}$$

where each p_k value for $k \leq k-1$ is normalized by the higher order term coefficient, p_k. From the possible roots, we pare back the results to only include real roots between $[0, 1]$. We know there will always be one root in this interval equal to 1 because of the nature of Eq. 5.6. However, if there is another root in the $[0, 1]$ interval, that is the root we focus on. Moving onto the second step in the process, z number of samples are produced from a normal distribution with $\mu = 0$ and $\sigma = 1$. This creates a matrix of z by k perturbations, Z, remembering that k is the length of the polynomial. Each z perturbation is applied to its respective coefficient. The Laub-Xia algorithm calls for a δ value to make the perturbations smaller, since it assesses the perturbation effects as they approach 0. Hence, we let $\delta = \sqrt{||x||2^{-16}}$.

Furthermore, for a single trial, once perturbations are applied, the root falling within $[0, 1]$, $\tilde{\mathbf{x}}$, of the perturbed polynomial, $\tilde{p}(x)$, is used in the component-wise division. This division is given by

$$\frac{|\tilde{\mathbf{x}} - \mathbf{x}|}{\delta|\mathbf{x}|}, \tag{5.10}$$

which finds the difference between the true root and the perturbed root, then nor-

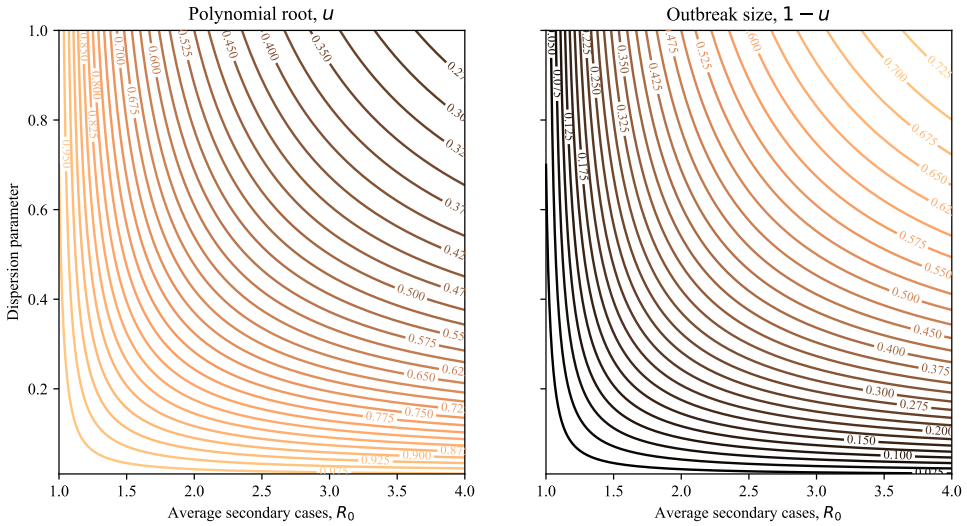

Figure 5.1: PGF root value and outbreak sizes: The contour lines of the PGF roots and outbreak size proportions are shown for average secondary cases,R_0 and the dispersion parameter, α, combinations. From the assumptions of Getz *et al.*, the polynomial root and the outbreak size plots are the inverse of each other [34]. A more heterogeneous network is represented by a lower dispersion value, and a more homogeneous network correlates to larger dispersion values.

malizes by δ and the true root.

5.3 Case Studies

To test the affect of added error on a given degree distribution, we construct a perturbed PGF. Remember, the error, δZ, is scaled by the coefficient, q_k, given by

$$\tilde{G}_1(x) = \sum_{k=0} q_k(1+\delta Z)x^k. \tag{5.11}$$

These coefficients are renormalized after inputting the error to preserve the nature of the PGF.

5.3.1 Negative binomial simulations

Negative binomial distributions commonly define distributions of secondary cases in the public health sector. We model the spread of disease as a random branching process [73], where an infected individuals generates a random number of individuals to infect, given by $G_0(x)$. Infection counts can either be tracked or understood through a giant component analysis, as detailed in Sec. 5.2.2. From the assumptions listed in Sec. 5.2.1, we define the terms of both $G_0(x)$ and $G_1(x)$ from a negative binomial distribution as,

$$q_k = \frac{\Gamma(k+\alpha)}{k!\Gamma(\alpha)}\Big(\frac{R_0}{R_0+\alpha}\Big)^k\Big(\frac{\alpha}{R_0+\alpha}\Big)^\alpha, \tag{5.12}$$

where R_0 is the average number of secondary cases, and α is the dispersion parameter. This distribution will then be perturbed according to Eq. (5.11). Before applying error, we derive the true polynomial root values from Eq. (5.6) and their respective final outbreak size, as seen in Fig. 5.1 to gauge each system's expected outcome.

5.3.2 Erdös-Rényi graph simulations

The above analysis considers transmission on the complete network, or that we use a percolated network directly, along with a variable dispersion value. To assess a distribution with a dispersion parameter, our second case focuses on an infinite Erdös-Rényi (ER) graph with a Poisson degree distribution. To consider percolation on the graph, we vary the occupation probability or transmission term in a disease dynamics

context, T. Before applying T, we first define each term in the Poisson distribution as

$$q_k = \frac{\lambda^k e^{-\lambda}}{k!}, \tag{5.13}$$

where λ is average degree. Similar to Sec. 5.3.1, we define $p_k = q_k$. Percolation is performed on this graph by removing edges from the graph uniformly at random. Each edge is removed with probability $1-T$ and remains in the graph with probability T. As seen in Sec. 3.3, the presence of each edge can then be modeled by a Bernoulli random variable with a probability generating function,

$$(1 - T) + Tx. \tag{5.14}$$

The resulting degree distribution of the percolated network is gained by composing the PGF of the un-percolated degree distribution with the PGF for a Bernoulli random variable,

$$G_1((1 - T) + Tx). \tag{5.15}$$

5.4 Results

5.4.1 Negative binomial: small outbreak sensitivity

The simulations described in Section 5.3.1 define perturbed systems on negative binomial degree distributions. The range of parameter values are $[1, 4]$ for R_0 and

$[0.01, 1]$ for α, allowing for both homogeneous and heterogeneous populations to be analyzed. For each combination of values in the parameter space, we calculate the condition number of all perturbed roots simulated. Figure 5.2 illustrates the condition number magnitudes of all perturbed roots. The darkest shades indicate the largest magnitudes, while the lightest shades correspond to the smallest magnitudes.

For ease of comparison, Fig. 5.3 overlays both the true polynomial roots and outbreak sizes over the condition numbers displayed in Fig. 5.2. We see the highest sensitivity following the contours of the largest roots, in the range of 0.925 to 0.950. This also means the highest sensitivities correlate to the smallest final outbreak proportions or 0.050 to 0.075. This conclusion that small final outbreak sizes are more sensitivity to noise follows naturally as the giant component could be eliminated from the system with a level of noise entirely. We see the polynomial root of 0.950 lies on the edge of the condition number decline. This means once a root value is smaller than 0.950 and greater than an outbreak size of 0.050, the system is much less susceptible to output variation from input perturbations. This part of the parameter space showcases a transition area that is important to be aware of for the confidence in final outbreak size predictions.

Another observation from Fig. 5.3 is that the parameter values of these large roots and small outbreaks occur across all R_0 values. For R_0 values between 1 and 1.25, the variation in the dispersion parameter does change the sensitivity, which indicates a larger sensitivity for more homogeneous systems with R_0 near 1. However, low dispersion values for the R_0 range of $[1.25, 1.5]$ also show sensitive systems. For larger values of R_0, specifically between 3 and 4, we see a higher magnitude than the center of the heat map. This indicates that the sensitivity of polynomial roots extends to larger

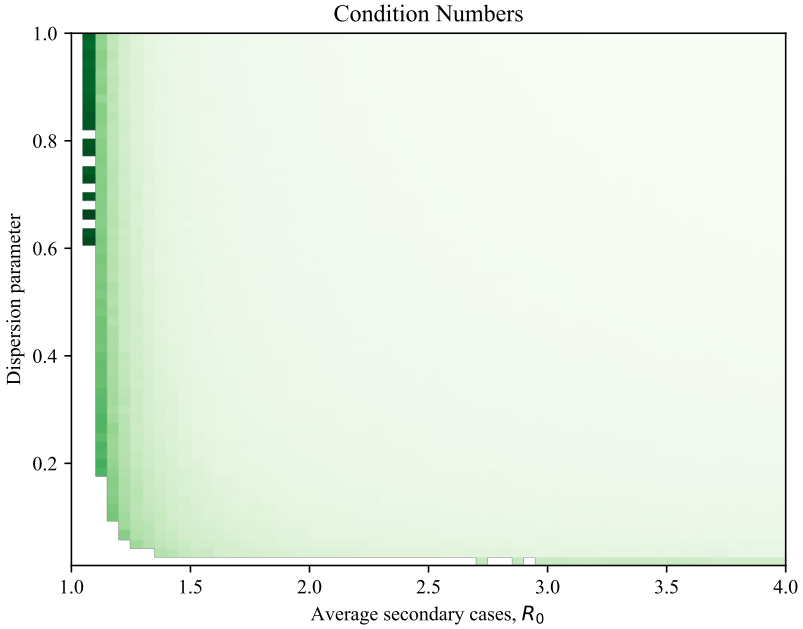

Figure 5.2: Negative binomial condition numbers: The heat map of condition numbers is shown for average secondary cases, R_0, and dispersion parameter, α, combinations. The darker parts of the map indicate larger condition numbers. We notice the most variable systems to be homogeneous systems with lower R_0 values. The white space at the lowest R_0 values are due the support of the distribution being large to simulate and infinite distribution support.

R_0 values. Thus, even with a large average secondary case value, the heterogeneity of the systems needs to be accounted for to properly assess the effect of noise.

5.4.2 Erdös-Rényi graphs: sensitive thresholds

For the perturbed ER graphs with varying T, as described in Sec. 5.3.2, we specify a range of T values from $[0.15, 0.9]$. These probabilities are applied to λ or the mean degree values, which range from $[1, 3]$. This changes the epidemic thresholds, as seen in the top panel of Fig. 5.4. For all T values and the mean degree values, we calculate

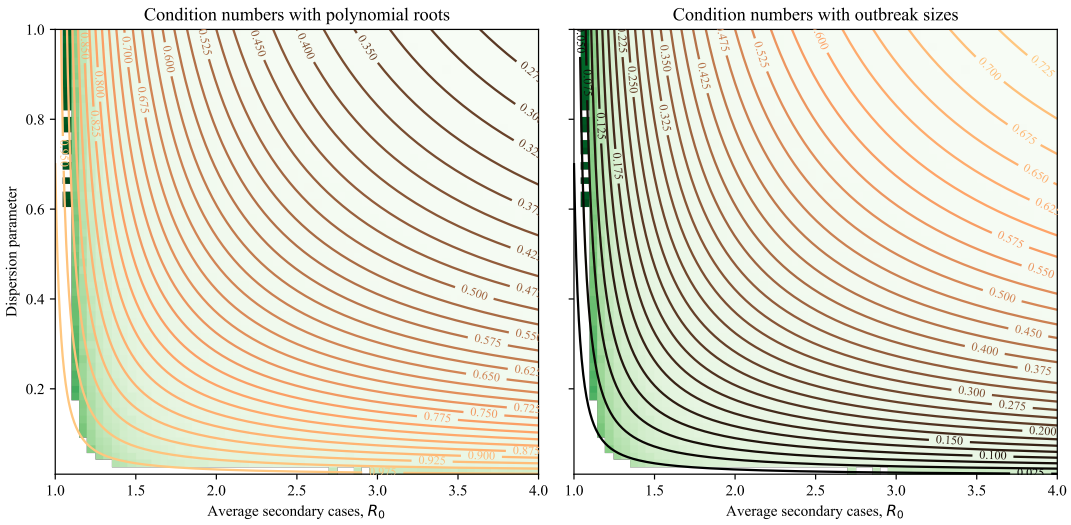

Figure 5.3: Negative binomial condition numbers overlaid with polynomial root and outbreak size: The heat map of condition numbers and variances are shown in both panels, however, the polynomial root and outbreak size are plotted on the same axes. The largest condition numbers, meaning the most sensitive values, correspond to the largest root values and smallest outbreak sizes. For systems make result in the smallest giant component or smallest outbreak appear to be most sensitive to perturbations.

the condition number, resulting in the bottom panel of Fig. 5.4. For each system, Fig. 5.4 illustrates a spike in the condition number at the critical transition when the giant component emerges. After each peak, the condition number decreases quickly at first, then the plots show a gradual decrease back to a lower sensitivity.

Similar to Sec. 5.4.1, the largest condition numbers occur at low giant component proportions, showcasing the sensitivity around giant component thresholds. Moreover, as the occupation probability changes, we expect the threshold values to increase, yet, we also see the condition numbers increase across T. The reason for the condition numbers being much larger for low occupation probabilities could come from the more gradual increase in outbreak size from 0, as seen from the line for $T = 0.30$ in the top panel of Fig. 5.4. Investigating this variation across condition

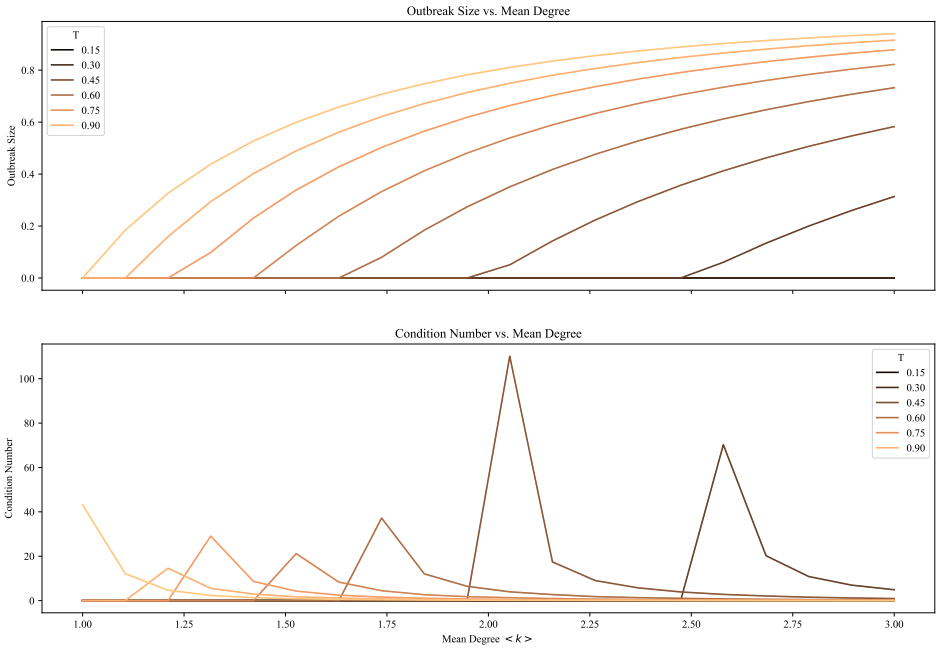

Figure 5.4: Percolated Erdös-Rényi: The condition numbers for ER graphs with occupation probability in the range of, $T = [0.15 - 0.90]$ are displayed. **Top**: The outbreak size versus the average degree of the network λ. **Bottom**: The condition number for each system corresponds to each outbreak size. It is clear that the condition number spikes at and around the critical transition when the giant component emerges.

numbers will be the subject of future work.

5.5 Discussion

Uncertainty on how perturbation effects modeling frameworks inhibit the accuracy of their outcomes and interpretation. When using PGFs for giant component analysis, it is necessary to know if the data being observed may yield varied results rather than one true result. We evaluated this framework for two cases, one specifying an

Disease	Year	R_0	k	Ref.
SARS	2003	1.63	0.16	[56, 54, 79]
Smallpox	1958-1973	3.19	0.37	[56, 57]
Influenza (Baltimore)	2009	1.77	0.94	[95, 32]
COVID-19	2020	2.5	0.1	[55, 27, 90, 93]

Table 5.1: Table of various diseases: This table provides specific cases of negative binomial parameters, which correlate to particular disease outbreaks.

epidemiological forecasting application, the other for a random graph. Both cases exhibit heightened sensitivity around the small outbreak or giant component sizes.

For infectious disease modeling that uses negative binomial distributions, there are many diseases that vary in heterogeneity, as seen in Table 5.1. From these different regimes, we can comment on their varying sensitivities to error and final outbreak proportions. For example, SARS may fall near a regime that has heighten sensitivity to noise. Therefore, with a system that has noise, there may not be a large outbreak predicted, when it truly could have one. Likewise for the ER graphs, extra analysis for varying condition values for small occupation probability is needed to further understand how perturbations affect this system.

In light of these results, this chapter aims future work not only towards the shown case studies, but also analyze the sensitivity of other branching process applications. As detailed in Sec. 3.3, there are other outcomes of PGFs that pose venues for sensitivity analyses like the work presented in this chapter. For example, how large and on which inputs do perturbations of the contact distribution from Chapter 4 affect the comparison metrics? Does perturbing the system in small but consistent increments change the metrics more than one large perturbation? These questions need to be evaluated for understand the scope of multiple PGF modeling tools.

As a general framework for branching process sensitivity evaluation, we recognize

the difficulty in seeing how random noise affects a probability distribution. However, regardless of the outcome of a PGF model, an altered probability distribution will warrant some change in an output. This analysis showcases that the regimes to be concerned about are those that produce outcomes close to 0. These regimes are tipping points for extinction and slow persistence. Also, the question of how much noise changes these sensitive regimes drastically is one to consider. This work hopes to present a better understanding of how to interpret results from PGFs, which account for the effects of imperfect data collection. These interpretations the allow modelers to more accurately convey results to the public.

Acknowledgments

M.C.B. is supported as a Fellow of the National Science Foundation under NRT award DGE-1735316. C.M.D., J.-G.Y. and L.H.-D. acknowledge financial support from the National Institutes of Health 1P20 GM125498-01 Centers of Biomedical Research Excellence Award.

Chapter 6

Conclusion

People desire a single 'correct' answer, prediction, or outcome from modeling tools. However, the dynamics of the world are messy and random. Nevertheless stochastic modeling encapsulates the unexpected trajectories in nature that modelers hope to capture. Hence, this body of work focuses on two probabilistic tools from the tool shed: master equations and probability generating functions. Each chapter of this thesis not only demonstrates expansions in methodology and new interpretations of these two modeling frameworks, but also evaluates subject specific uncertainties for other scientists, decision makers, and modelers.

Firstly, Chapter 2 presents a within-host HPV infection progression model, that leverages master equation moments for faster computations. From moment information, we inform uncertainties for temporal probability of extinction, along with temporal persistence and viral load information. With the ability to test multiple tissue structures, the consideration of an individual's age changes the viral load progression outcomes. This framework has the potential to be extended to other cell specific infections that affect various parts of the body, along with informing population-level

model parameters.

Next, Chapter 4 extends a temporal PGF analysis, which defines a temporal transmission term, affected by interventions on the system. During this process, we notice the difficulty in comparing the cumulative case count probability distributions for differing intervention strategies. Consequently, we define four comparison metrics to aid disease spread mitigation decisions. These comparison metrics showcase the effectiveness of certain interventions over others, along with making model outcomes clearer for decision makers.

Finally, Chapter 5 delves into the first sensitivity analysis for a PGF giant component analysis. For two cases and their parameter spaces, a negative binomial distributions and an Erdös-Rényi random graph, a condition number analysis evaluates the sensitivity of final outbreak or giant component size. The highest sensitivity occurred at the critical threshold value for each case. From this result, modelers can gauge the effects of errors on this model outcome. This work also establishes a framework for sensitivity analyses on other PGF outcomes.

6.1 Methodological assumptions

In the midst of showcasing three models that push the edges of knowledge for the mathematical modeling of disease dynamics, there is another important common thread to discuss. In each chapter, the methodological sections began with an Assumptions sub-section. This sub-section is inspired from Mitchell *et al.*'s work on model cards [65]. Model cards exist in the machine learning side of the modeling tool shed, detailing extensive information on the model and its creators. This includes the

details on the training data, performance measures, ethical considerations, intended use, and much more. While all mathematical models of disease do not take in the same data, parameters, or give comparable results, clear assumptions on the model and its intent are essential aspects to convey to readers. As George Box said "All models are wrong, but some are useful" [18], so it is imperative to detail explicit assumptions and intentions to distinguish the useful models.

It is the hope of this work to encourage others to state information inspired by the model card framework and display the information explicitly. For example, a statement on the modelers and their backgrounds provides insights as to why certain models choices are made. Other valuable information for readers to be aware of are the sensitivities of a model, with the results of Chapter 5 being an example of this. This type of information makes readers aware of the caveats and intentions long before reading the discussion, and absorbing the methodologies and take-away points. The range and uses of the tools in the tool shed need to be clear, otherwise someone could try to hammer a screw into a board.

6.2 Parameter literature review

In the processes of writing Chapter 2, an extensive literature review was conducted for the cellular division dynamics and HPV viral load output. One of the sources that informed the cellular dynamics of the model details rate defaults and ranges for both data-derived and in vivo, cell culture experiment, literature estimates. This source compared the data-derived and literature estimates, some of the defaults and ranges were quite different, but were not expected to be identical. For example, the

data-derived parabasal cell replication rate average was 0.0082, with no range given. Whereas the literature estimate average given was 0.39, with a range of [0.2, 1] [70]. The question of which parameter choice to use is up to the reader, making it unclear what the benefits and downsides of each estimate are.

In addition, a reference for the in-vivo experiment is given, leading the reader to explore a different publication. Upon reading this publication, it was difficult to understand how the original ranges were derived. The struggle to understand parameter values through references and reported data showcase areas of improvement. It is with this experience that I acknowledge an opportunity for modelers and other scientists to consider making all data from their findings available and interpretable to readers. Without access to the data informing an estimated parameter value, it is hard to account for parameter variation. This variation, or lack there of, aids in accommodating the models we construct to inform unknown outcomes.

6.3 Future multi-scale model

From the frameworks of Chapters 2 and 4, we can define a case study for a multi-scale model framework. Considering HPV again, we take the information from Chapter 2 to define distributions describing the model outcomes. Those being viral load, extinction probability, and distribution of dead cells from persistent infection of a randomly selected infected individual in a population. The viral load correlates to the rate of transmitting HPV, β in the Chapter 4 model, to a contact in the contact degree distribution. One way to determine this is using the average value from the viral load zero-inflated distribution for all β values. The second is to pull a β value

from the zero-inflated distribution determined by the first and second moment of the dead cells to accommodate heterogeneity in the probability of infection over time. Depending on the age of an individual in the population, this would dictate the cell-type structure moments used. Moving onto the recovery rate, γ, this correlates directly to the probability of extinction. We can define γ by number of people in a potential population multiplied by the extinction probability over time. Finally, if an individual has been has not recovered past a certain time, say the time that correlates with a probability of extinction of 50%, then the moments for the dead cells of the persistent infections will impose will be interpreted to viral load for a new distribution for β. Similarly, the age of an individual will inform the system structure and respective moments.

After defining the rates of infecting another individual or a recovery, the last input of the temporal and probabilistic model of Chapter 4 is defining the contact distribution. From the literature, skewed distributions such as power-law or Weibull distributions have been used by modelers to define a generalized sexual contact network [41, 25, 28]. With a temporal sexual contact network, we could define the possibility of re-infection by allowing contact to occur again. Thus, the model can be run with or without an HPV vaccine implementation. From demographic information for a modeled population, the comparison metrics could inform the likelihood of persistence and cancer risk in a population when evaluating the new case distribution over time.

6.4 Final thoughts

While the work of this thesis is far from presenting the perfect methods to uncover uncertainties in unknown with-in host dynamics, probability distribution comparisons, or sensitivity analyses for PGFs, it is a start. A start that I am eager to discuss with anyone interested, as I have been thinking about the body of this work for almost five years. I thank the reader for taking the time and space to get this far, and leave you with the visual contributions of this thesis.

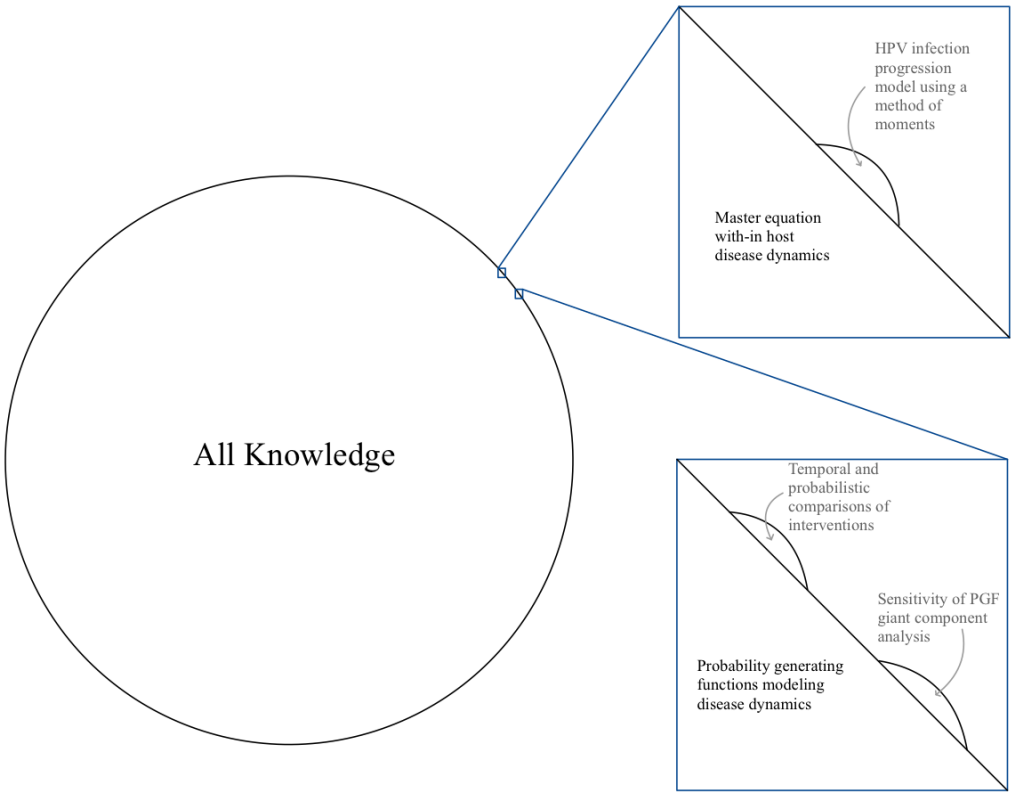

Figure 6.1: Visual contributions: This figure is inspired by "The illustrated guide to a Ph.D." by Matt Might, which I read when considering a higher education [61].

Bibliography

[1] A. Aleta et al. "A data-driven assessment of early travel restrictions related to the spreading of the novel COVID-19 within mainland China". In: *Chaos Solitons Fractals* 139 (2020), p. 110068. DOI: 10.1016/j.chaos.2020.110068.

[2] A. Allen. *andrea-allen/epintervene: Initial Release*. Version v1.0.0. July 2021. DOI: 10.5281/zenodo.5076514. URL: https://doi.org/10.5281/zenodo.5076514.

[3] A. Allen. *andrea-allen/pgf-networks*. URL: https://github.com/andrea-allen/pgf-networks.

[4] A. J. Allen et al. "Predicting the diversity of early epidemic spread on networks". In: *Phys. Rev. Res.* 4.1 (2022), p. 013123.

[5] L. J. S. Allen. "A primer on stochastic epidemic models: Formulation, numerical simulation, and analysis". In: *Infect. Dis. Model.* 2.2 (2017), pp. 128–142.

[6] L. J. S. Allen. "Stochastic population and epidemic models". In: *Math. Biosci.* 1 (2015), pp. 120–128.

[7] B. M. Althouse et al. *The unintended consequences of inconsistent pandemic control policies*. Preprint medRxiv:2020.08.21.20179473. 2020. DOI: 10.1101/2020.08.21.20179473.

[8] T. S. N. Asih et al. "The dynamics of HPV infection and cervical cancer cells". In: *B. Math. Biol.* 78 (2016), pp. 4–20.

[9] N. Bacaër. "Daniel Bernoulli, d'Alembert and the inoculation of smallpox (1760)". In: *A Short History of Mathematical Population Dynamics*. London: Springer London, 2011, pp. 21–30. ISBN: 978-0-85729-115-8. DOI: 10.1007/978-0-85729-115-8_4. URL: https://doi.org/10.1007/978-0-85729-115-8_4.

[10] F. Ball and P. Donnelly. “Strong approximations for epidemic models”. In: *Stoch. Proc. Appl.* 55.1 (1995), pp. 1–21. ISSN: 0304-4149. DOI: https://doi.org/10.1016/0304-4149(94)00034-Q. URL: https://www.sciencedirect.com/science/article/pii/030441499400034Q.

[11] F. Ball, D. Mollison, and G. Scalia-Tomba. “Epidemics with two levels of mixing”. In: *Ann. Appl. Probab.* 7.1 (1997), pp. 46–89. DOI: 10.1214/aoap/1034625252. URL: https://doi.org/10.1214/aoap/1034625252.

[12] T. Beneteau et al. “Episome partitioning and symmetric cell divisions: Quantifying the role of random events in the persistence of HPV infections”. In: *PLoS Comput. Biol.* 17.9 (2021), e1009352.

[13] D. Bernoulli and D. Chapelle. “Essai d’une nouvelle analyse de la mortalité causée par la petite vérole, et des avantages de l’inoculation pour la prévenir”. In: (2023).

[14] P. Block et al. “Social network-based distancing strategies to flatten the COVID-19 curve in a post-lockdown world”. In: *Nat. Hum. Behav.* 4 (2020), pp. 588–596. DOI: 10.1038/s41562-020-0898-6.

[15] J. A. Bondy, U. S. R. Murty, et al. *Graph theory with applications*. Vol. 290. Macmillan London, 1976.

[16] M. C. Boudreau, J. A. Cohen, and L. Hébert-Dufresne. “Within-host infection dynamics with master equations and the method of moments: A case study of human papillomavirus in the epithelium”. In: *Manuscript currently under prepartion* (2024).

[17] M. C. Boudreau et al. “Sensitivity analysis of stochastic polynomial roots, and its application to epidemic forecasting and random graphs”. In: *Manuscript currently under prepartion* (2024).

[18] G. E. P. Box. “Robustness in the strategy of scientific model building”. In: *Robustness in statistics*. Elsevier, 1979, pp. 201–236.

[19] W. E. Boyce and R. C. DiPrima. *Elementary differential equations and boundary value problems*. Wiley, 2020.

[20] D. K. Chu et al. “Physical distancing, face masks, and eye protection to prevent person-to-person transmission of SARS-CoV-2 and COVID-19: a systematic review and meta-analysis”. In: *Lancet* 395.10242 (2020), pp. 1973–1987.

[21] T. Churches and L. Jorm. “Flexible, Freely Available Stochastic Individual Contact Model for Exploring COVID-19 Intervention and Control Strategies: Development and Simulation”. In: *JMIR Public Health Surveill.* 6 (2020), e18965. DOI: 10.2196/18965.

[22] S. M. Ciupe and J. M. Heffernan. “In-host modeling”. In: *Infect. Dis. Model.* 2.2 (2017), pp. 188–202.

[23] E. Clayton et al. “A single type of progenitor cell maintains normal epidermis”. In: *Nature* 446.7132 (2007), pp. 185–189.

[24] N. G. Davies et al. “Effects of non-pharmaceutical interventions on COVID-19 cases, deaths, and demand for hospital services in the UK: a modelling study”. In: *Lancet Public Health* 5 (2020), e375–e385. DOI: 10.1016/S2468-2667(20)30133-X.

[25] B. F. De Blasio, Å. Svensson, and F. Liljeros. “Preferential attachment in sexual networks”. In: *PNAS* 104.26 (2007), pp. 10762–10767.

[26] J. Doorbar. “Papillomavirus life cycle organization and biomarker selection”. In: *Dis. Markers* 23.4 (2007), pp. 297–313.

[27] A. Endo et al. “Estimating the overdispersion in COVID-19 transmission using outbreak sizes outside China”. In: *Wellcome open research* 5 (2020).

[28] A. Endo et al. “Heavy-tailed sexual contact networks and monkeypox epidemiology in the global outbreak, 2022”. In: *Science* 378.6615 (2022), pp. 90–94.

[29] J. M. Epstein. “Why model?” In: *JASSS* 11.4 (2008), p. 12.

[30] L. Euler. “Solutio problematis ad geometriam situs pertinentis”. In: *Commentarii academiae scientiarum Petropolitanae* (1741), pp. 128–140.

[31] B. K. Fosdick et al. “Configuring random graph models with fixed degree sequences”. In: *SIAM Rev.* 60.2 (2018), pp. 315–355.

[32] C. Fraser et al. “Influenza transmission in households during the 1918 pandemic”. In: *Am. J. Epidemiol.* 174.5 (2011), pp. 505–514.

[33] P. A. Garner-Hamrick and C. Fisher. “HPV episomal copy number closely correlates with cell size in keratinocyte monolayer cultures”. In: *Virology* 301.2 (2002), pp. 334–341.

[34] W. Getz and J. Lloyd-Smith. “Basic methods for modeling the invasion and spread of contagious disease”. In: *Disease evolution: models, concepts, and data analyses* 71 (Jan. 2005).

[35] E. N. Gilbert. “Random Graphs”. In: *Ann. Math. Statist.* 30(4) (1959), pp. 1141–1144. DOI: doi:10.1214/aoms/1177706098.

[36] D. T. Gillespie. “Stochastic simulation of chemical kinetics”. In: *Annu. Rev. Phys. Chem.* 58 (2007), pp. 35–55.

[37] D.T. Gillespie. “Exact stochastic simulation of coupled chemical reactions”. In: *J. Phys. Chem.* 81.25 (1977), pp. 2340–2361.

[38] P. E. Gravitt et al. “The known unknowns of HPV natural history”. In: *J. Clin. Invest.* 121.12 (2011), pp. 4593–4599.

[39] G. Haag. “Modelling with the Master equation”. In: *Solution Methods* (2017).

[40] P. Haccou, P. Jagers, and V. A. Vatutin. *Branching processes: variation, growth, and extinction of populations*. 5. Cambridge University Press, 2005.

[41] D. T. Hamilton, M. S. Handcock, and M. Morris. “Degree distributions in sexual networks: a framework for evaluating evidence”. In: *Sex. Transm. Dis.* 35.1 (2008), pp. 30–40.

[42] T. E. Harris et al. *The theory of branching processes*. Vol. 6. Springer Berlin, 1963.

[43] L. Hébert-Dufresne and B. M. Althouse. “Complex dynamics of synergistic coinfections on realistically clustered networks”. In: *PNAS* 112.33 (2015), pp. 10551–10556.

[44] L. Hébert-Dufresne, O. Patterson-Lomba, and B. M. Goerg G. M .and Althouse. “Pathogen mutation modeled by competition between site and bond percolation”. In: *Phys. Rev. Lett.* 110.10 (2013), p. 108103.

[45] H. W. Hethcote. “Three basic epidemiological models”. In: *Applied mathematical ecology*. Springer, 1989, pp. 119–144.

[46] E. Kenah and J. M. Robins. “Second look at the spread of epidemics on networks”. In: *Phys. Rev. E.* 76.3 (2007), p. 036113.

[47] D. G. Kendall. “On the generalized "birth-and-death" process”. In: *Ann. Math. Stat.* 19.1 (1948), pp. 1–15.

[48] W. O. Kermack and A. G. McKendrick. “A contribution to the mathematical theory of epidemics-I.” In: *Proc. R. Soc. Lond. A. Math. Phys. Sci.* 115.772 (1927), pp. 700–721.

[49] W. O. Kermack and A. G. McKendrick. “Contributions to the mathematical theory of epidemics- III. Further studies of the problem of endemicity”. In: *Proc. R. Soc. Lond. A. Math. Phys. Sci.* 141.843 (1933), pp. 94–122.

[50] W. O. Kermack and A. G. McKendrick. “Contributions to the mathematical theory of epidemics-II. The problem of endemicity”. In: *Proc. R. Soc. Lond. A. Math. Phys. Sci.* 138.834 (1932), pp. 55–83.

[51] S. Kojaku et al. “The effectiveness of backward contact tracing in networks”. In: *Nat. Phys.* 17 (2021), pp. 652–658. DOI: 10.1038/s41567-021-01187-2.

[52] A. J. Laub and J. Xia. “Statistical condition estimation for the roots of polynomials”. In: *SIAM J. Sci. Comput.* 31.1 (2008), pp. 624–643.

[53] R. J. Lempert, S. W. Popper, and S. C. Bankes. “Robust Decision Making: Coping with Uncertainty”. In: *Futurist* 44 (2010), pp. 47–48. URL: https://www.proquest.com/openview/624dcdc33173abd5dff56acf88baf624/1.

[54] G. M. Leung et al. “Seroprevalence of IgG antibodies to SARS-coronavirus in asymptomatic or subclinical population groups”. In: *Epidemiol. Infect.* 134.2 (2006), pp. 211–221.

[55] Q. Li et al. “Early transmission dynamics in Wuhan, China, of novel coronavirus–infected pneumonia”. In: *N. Engl. J. Med.* (2020).

[56] J. O. Lloyd-Smith et al. “Superspreading and the effect of individual variation on disease emergence”. In: *Nature* 438.7066 (2005), pp. 355–359.

[57] T. M. Mack et al. “Epidemiology of smallpox in West Pakistan: I. Acquired immunity and the distribution of disease”. In: *Am. J. Epidemiol.* 95.2 (1972), pp. 157–168.

[58] E. Meites et al. “Human Papilloma Virüs”. In: *Centers for disease control and prevention (CDC).[Erişim tarihi: October 2020]. Erişim adresi: Pinkbook/ HPV/ Epidemiology of Vaccine Preventable Diseases/ CDC* (2020).

[59] L. A. Meyers. “Contact network epidemiology: Bond percolation applied to infectious disease prediction and control”. In: *B. Am. Math. Soc.* 44.1 (2007), pp. 63–86.

[60] L. A. Meyers, M. E. J. Newman, and B. Pourbohloul. “Predicting epidemics on directed contact networks”. In: *J. Theor. Biol.* 240.3 (2006), pp. 400–418.

[61] M. Might. “The illustrated guide to a Ph. D”. In: *Texte librement adapté, Images sous licence Creative Commons Attribution-NonCommercial* (2010).

[62] J. C. Miller. “A primer on the use of probability generating functions in infectious disease modeling”. In: *Infect. Dis. Model.* 3 (2018), pp. 192–248.

[63] J. C. Miller. “Percolation and epidemics in random clustered networks”. In: *Phys. Rev. E.* 80.2 (2009), p. 020901.

[64] G. J. Milne et al. “A Small Community Model for the Transmission of Infectious Diseases: Comparison of School Closure as an Intervention in Individual-Based Models of an Influenza Pandemic”. In: *PLOS One* 3 (2008), e4005. DOI: 10.1371/journal.pone.0004005.

[65] M. Mitchell et al. “Model cards for model reporting”. In: *Proceedings of the conference on fairness, accountability, and transparency.* 2019, pp. 220–229.

[66] C. A. Moody. “Mechanisms by which HPV induces a replication competent environment in differentiating keratinocytes”. In: *Viruses* 9.9 (2017), p. 261.

[67] C. Moore and M. E. J. Newman. “Exact solution of site and bond percolation on small-world networks”. In: *Phys. Rev. E.* 62.5 (2000), p. 7059.

[68] M. D. Morris. “Factorial sampling plans for preliminary computational experiments”. In: *Technometrics* 33.2 (1991), pp. 161–174.

[69] C. L. Murall, C. T. Bauch, and T. Day. “Could the human papillomavirus vaccines drive virulence evolution?” In: *P. R. Soc. B-Biol. Sci.* 282.1798 (2015), p. 20141069.

[70] C. L. Murall et al. “Epithelial stratification shapes infection dynamics”. In: *PLoS Comput. Biol.* 15.1 (2019), e1006646.

[71] M. E. J. Newman. *Networks*. Oxford university press, 2018.

[72] M. E. J. Newman. “Spread of epidemic disease on networks”. In: *Phys. Rev. E.* 66.1 (2002), p. 016128.

[73] M. E. J. Newman, S. H. Strogatz, and D. J. Watts. “Random graphs with arbitrary degree distributions and their applications”. In: *Phys. Rev. E.* 64.2 (2001), p. 026118.

[74] P.-A. Noël et al. “Time evolution of epidemic disease on finite and infinite networks”. In: *Phys. Rev. E.* 79.2 (2009), p. 026101.

[75] K. Ohno et al. “A computational model of the epidermis with the deformable dermis and its application to skin diseases”. In: *Sci. Rep-UK* 11.1 (2021), p. 13234.

[76] R. Pastor-Satorras and A. Vespignani. “Immunization of complex networks”. In: *Phys. Rev. E* 65 (2002), p. 036104. DOI: 10.1103/PhysRevE.65.036104.

[77] C. M. Peak et al. “Comparing nonpharmaceutical interventions for containing emerging epidemics”. In: *PNAS* 114 (2017), pp. 4023–4028. DOI: 10.1073/pnas.1616438114.

[78] W. Prendiville and R. Sankaranarayanan. *Colposcopy and treatment of cervical precancer*. International Agency for Research on Cancer, World Health Organization, 2017.

[79] S. R. Quah and L. Hin-Peng. “Crisis prevention and management during SARS outbreak, Singapore”. In: *Emer. Infect. Dis.* 10.2 (2004), p. 364.

[80] S. F. Rosenblatt et al. “Immunization strategies in networks with missing data”. In: *PLOS Comput. Biol.* 16 (2020), e1007897. DOI: 10.1371/journal.pcbi.1007897.

[81] S. M. Ross. *Stochastic processes*. John Wiley & Sons, 1995.

[82] M. D. Ryser, P. E. Gravitt, and E. R. Myers. “Mechanistic mathematical models: An underused platform for HPV research”. In: *Papillomavirus Res.* 3 (2017), pp. 46–49.

[83] M. D. Ryser, E. R. Myers, and R. Durrett. “HPV clearance and the neglected role of stochasticity”. In: *PLoS Comput. Biol.* 11.3 (2015), e1004113.

[84] A. Saltelli et al. *Global sensitivity analysis: the primer*. John Wiley & Sons, 2008.

[85] E. R. Sari, F. Adi-Kusumo, and L. Aryati. “Mathematical analysis of a SIPC age-structured model of cervical cancer”. In: *Mathematical Biosciences and Engineering* 19.6 (2022), pp. 6013–6039.

[86] J. C. Sierra-Rojas et al. “Modeling and Mathematical Analysis of the Dynamics of HPV in Cervical Epithelial Cells: Transient, Acute, Latency, and Chronic Infections”. In: *Comput. Math. Method. M.* 2022 (2022).

[87] E. Speranza. “Understanding virus–host interactions in tissues”. In: *Nat. Microbiol.* 8.8 (2023), pp. 1397–1407.

[88] M. A. Stanley. “Epithelial cell responses to infection with human papillomavirus”. In: *Clin. Microbiol. Rev.* 25.2 (2012), pp. 215–222.

[89] D. Stauffer and A. Aharony. *Introduction to percolation theory*. CRC press, 2018.

[90] H. Streeck et al. “Preliminary result and conclusions of the COVID-19 case cluster study (Gangelt Municipality)”. In: (2020).

[91] R. M. Stuart et al. “HPVsim: An agent-based model of HPV transmission and cervical disease”. In: *PLoS Comput. Biol.* 20.7 (2024), e1012181.

[92] C. Sun and Z. Zhai. “The efficacy of social distance and ventilation effectiveness in preventing COVID-19 transmission”. In: *Sustain. Cities Soc.* 62 (2020), p. 102390. DOI: 10.1016/j.scs.2020.102390.

[93] D. Sutton et al. “Universal screening for SARS-CoV-2 in women admitted for delivery”. In: *N. Engl. J. Med.* 382.22 (2020), pp. 2163–2164.

[94] D.C. Swan et al. “Human papillomavirus (HPV) DNA copy number is dependent on grade of cervical disease and HPV type”. In: *J. Clin. Microbiol.* 37.4 (1999), pp. 1030–1034.

[95] J. K. Taubenberger and D. M. Morens. “1918 Influenza: the mother of all pandemics”. In: *Revista Biomedica* 17.1 (2006), pp. 69–79.

[96] T. G. Vaughan, P. D. Drummond, and A. J. Drummond. “Within-host demographic fluctuations and correlations in early retroviral infection”. In: *J. Theor. Biol.* 295 (2012), pp. 86–99.

[97] L. Wessel et al. “Public health interventions for epidemics: implications for multiple infection waves”. In: *BMC Public Health* 11 (2011), S2. DOI: 10.1186/1471-2458-11-S1-S2.

[98] E. R. White and L. Hébert-Dufresne. “State-level variation of initial COVID-19 dynamics in the United States”. In: *PLOS One* 15 (2020), e0240648. DOI: 10.1371/journal.pone.0240648.

[99] H. S. Wilf. *generatingfunctionology*. CRC press, 2005.

[100] J. Wu et al. “Sensitivity analysis of infectious disease models: methods, advances and their application”. In: *J. R. Soc. Interface* 10.86 (2013), p. 20121018.

Appendix

Three-cell-type moment equations

The moment equations for Sec. 2.3.1 are given by:

$$
\begin{aligned}
\frac{d}{dt}\langle b\rangle &= (\beta-\delta)\langle b\rangle \\
\frac{d}{dt}\langle p\rangle &= (\rho-\theta)\langle p\rangle + (2\delta+\gamma)\langle b\rangle \\
\frac{d}{dt}\langle d\rangle &= \theta\langle p\rangle \\
\frac{d}{dt}\langle b^2\rangle &= 2(\beta-\delta)\langle b^2\rangle + (\beta+\delta)\langle b\rangle \\
\frac{d}{dt}\langle p^2\rangle &= (\theta+\rho)\langle p\rangle + (2\rho-2\theta)\langle p^2\rangle + (\gamma+4\delta)\langle b\rangle + (2\gamma+4\delta)\langle bp\rangle \\
\frac{d}{dt}\langle d^2\rangle &= \theta\langle p\rangle + 2\theta\langle pd\rangle \\
\frac{d}{dt}\langle bp\rangle &= (\beta-\theta+\rho-\delta)\langle bp\rangle + (\gamma+2\delta)\langle b^2\rangle - 2\delta\langle b\rangle \\
\frac{d}{dt}\langle pd\rangle &= (\rho-\theta)\langle pd\rangle + \theta\langle p^2\rangle - \theta\langle p\rangle + (\gamma+2\delta)\langle bd\rangle \\
\frac{d}{dt}\langle bd\rangle &= \beta\langle b\rangle + \theta\langle bp\rangle - \delta\langle bd\rangle.
\end{aligned}
$$

Five-cell-type moment equations

The moment equations for Sec. 2.3.2 are given by:

$$
\begin{aligned}
\frac{d}{dt}\langle b\rangle &= (\beta-\delta)\langle b\rangle \\
\frac{d}{dt}\langle p\rangle &= (\rho-\alpha)\langle p\rangle + (\gamma+2\delta)\langle b\rangle \\
\frac{d}{dt}\langle i\rangle &= \alpha\langle p\rangle - \sigma\langle i\rangle \\
\frac{d}{dt}\langle s\rangle &= \sigma\langle i\rangle - \theta\langle s\rangle \\
\frac{d}{dt}\langle d\rangle &= \theta\langle s\rangle \\
\frac{d}{dt}\langle b^2\rangle &= 2(\beta-\delta)\langle b^2\rangle + (\beta+\delta)\langle b\rangle \\
\frac{d}{dt}\langle p^2\rangle &= (\rho+\alpha)\langle p\rangle + (2\rho-2\alpha)\langle p^2\rangle + (2\gamma+4\delta)\langle bp\rangle + (\gamma+4\delta)\langle b\rangle \\
\frac{d}{dt}\langle i^2\rangle &= \alpha\langle p\rangle + 2\alpha\langle pi\rangle - 2\sigma\langle i^2\rangle + \sigma\langle i\rangle \\
\frac{d}{dt}\langle s^2\rangle &= \sigma\langle i\rangle + 2\sigma\langle is\rangle - 2\theta\langle s^2\rangle + \theta\langle s\rangle \\
\frac{d}{dt}\langle d^2\rangle &= 2\theta\langle sd\rangle + \theta\langle s\rangle \\
\frac{d}{dt}\langle bp\rangle &= (\beta+\rho-\delta-\alpha)\langle bp\rangle + (\gamma+2\delta)\langle b^2\rangle - 2\delta\langle b\rangle \\
\frac{d}{dt}\langle bi\rangle &= (\beta-\delta-\sigma)\langle bi\rangle + \alpha\langle bp\rangle \\
\frac{d}{dt}\langle bs\rangle &= (\beta-\delta-\theta)\langle bs\rangle + \sigma\langle bi\rangle \\
\frac{d}{dt}\langle bd\rangle &= (\beta-\delta)\langle bd\rangle + \theta\langle bs\rangle \\
\frac{d}{dt}\langle pi\rangle &= (\rho-\alpha-\sigma)\langle pi\rangle + (\gamma+2\delta)\langle bi\rangle + \alpha\langle p^2\rangle - \alpha\langle p\rangle \\
\frac{d}{dt}\langle ps\rangle &= (\rho-\theta-\alpha)\langle ps\rangle + (\gamma+2\delta)\langle bs\rangle + \sigma\langle pi\rangle
\end{aligned}
$$

$$\frac{d}{dt}\langle pd\rangle = (\rho-\alpha)\langle pd\rangle + (\gamma+2\delta)\langle bd\rangle + \theta\langle ps\rangle$$
$$\frac{d}{dt}\langle is\rangle = (-\theta-\sigma)\langle is\rangle + \alpha\langle ps\rangle + \sigma\langle i^2\rangle - \sigma\langle i\rangle$$
$$\frac{d}{dt}\langle id\rangle = \theta\langle is\rangle - \sigma\langle id\rangle + \alpha\langle pd\rangle$$
$$\frac{d}{dt}\langle id\rangle = \theta\langle is\rangle - \sigma\langle id\rangle + \alpha\langle pd\rangle$$
$$\frac{d}{dt}\langle sd\rangle = \theta\langle s^2\rangle - \theta\langle sd\rangle - \theta\langle s\rangle + \sigma\langle id\rangle$$