Title

Topics in Network Science: Modeling of Microbiome Populations in Interacting Hosts and an Application of Persistent Homology to Resource Coverage

Author

Johnson, Michael

Publication Date

2025

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA
Los Angeles

Topics in Network Science: Modeling of Microbiome Populations in Interacting Hosts and an Application of Persistent Homology to Resource Coverage

A dissertation submitted in partial satisfaction
of the requirements for the degree
Doctor of Philosophy in Mathematics

by

Michael Christopher Johnson

2025

ABSTRACT OF THE DISSERTATION

Topics in Network Science:
Modeling of Microbiome Populations in
Interacting Hosts and an Application of
Persistent Homology to Resource Coverage

by

Michael Christopher Johnson
Doctor of Philosophy in Mathematics
University of California, Los Angeles, 2025
Professor Mason Alexander Porter, Chair

In many scientific disciplines, it is important to study networks of interconnected agents. In this dissertation, we discuss two research projects in network science.

In the first project, we introduce a novel modeling framework for the microbiome dynamics of interacting hosts. The microbiomes of humans and animals play a critical role in their functioning and health. Researchers have studied the dynamics of microbiomes using models in the form of dynamical systems. Many classical ecological models are suitable only for modeling the microbiome of one isolated host. However, there is strong evidence that interactions between hosts significantly impact their microbiome compositions. In the study of microbiome dynamics, researchers employ metacommunity-theory models to investigate the effects of multiple-scale interactions. These models commonly assume a continuous dispersal of microbes between interacting hosts. However, many living hosts (such as humans) do not interact continuously and thus do not sustain a continuous dispersal of microbes. In this

dissertation, we introduce a novel modeling framework that considers the discrete nature of host interactions by using two parameters to separately control the interaction frequencies between hosts and the amount of microbiome exchange during each interaction. We derive analytical approximations for our modeling framework in three different regimes, and we compare the resulting approximate dynamics to simulations of our modeling framework for an illustrative model.

In the second project, we use persistent homology (PH) to identify holes in resource coverage. The geographical distribution of resources such as polling sites (i.e., locations where people vote), hospitals, COVID-19 vaccination sites, Department of Motor Vehicles (DMV) locations, and Planned Parenthood clinics is a significant factor in the equitability of access to those resources. Consequently, given the locations of a set of resource sites, it is important to quantify their geographical coverage and to identify underserved geographical regions. The information from PH allows us to infer holes in the distribution of polling sites. In our PH approach, we construct a distance function d that is based on the travel times to a resource. This distance function represents the costs of accessing a resource better than geographical distance. We apply our methodology to a case study of polling-site access in the 2016 United States presidential election. We analyze and compare the coverage of polling sites in Los Angeles County and five cities (Atlanta, Chicago, Jacksonville, New York City, and Salt Lake City).

The first project employs a theory-driven approach, and the second project employs a data-driven approach, yielding distinct insights into their respective applications.

The dissertation of Michael Christopher Johnson is approved.

Christopher A. Klausmeier

Christopher R. Anderson

Deanna M. Needell

Mason Alexander Porter, Committee Chair

University of California, Los Angeles

2025

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

Completing this PhD program has been the most significant and challenging achievement of my life. This would not have been possible without the professional and personal support of many people in my life.

Mason Porter is a mentor who has checked both of those boxes. Mason, I am so grateful that you were my advisor. Professionally, you have offered me many project opportunities, but more importantly you have provided me with constant support. The time that you take to meet with me (and the many other people that you advise) every week is uniquely generous. These meetings demonstrate your genuine care for those around you, and they were invaluable during my time in this program. I knew that every time I stumbled in the research process, I had at most a week before a meeting with you would help set me back on the right track. Our interactions have also been very personally important to me. Graduate school has been the most challenging thing that I have attempted in my life. My experience with research has been difficult at times. It has sometimes felt like I have spent weeks working on a project angle that ultimately does not work. I know now that this is a necessary part of the process, but for my first couple of years, these stalls in progress were extremely discouraging. I was very anxious to attend some of our early meetings and tell you there had not been tangible progress in the last week, or two, or three. Every time, you responded with care and support. That support was one of the key factors in helping me overcome the personal obstacles that I faced during this program. Thank you, Mason.

I also want to thank the many members of Mason's group I have had the pleasure to interact with. Listening to group members present at Networks Journal Club every week was extremely valuable as I was choosing research projects. More recently, I've appreciated all of your feedback and support when I was presenting my work at those meetings. I hope I don't miss anyone, but thank you David Beers, Weiqi Chu, Christine Craib, Theodore Faust, Linnéa Gyllingberg, Casey Johnson, Abigail Hickok, Leah Keating, Grace Li, Jiajie

Luo, Tung Nguyen, James Raj, Filippo Riscica Lizzio, Kye Shi, Sarah Tymochko, and Mia Zender.

I want to thank the rest of my dissertation committee—Christopher Anderson, Christopher Klausmeier, and Deanna Needell—for their guidance throughout this process.

I acknowledge the National Science Foundation for the financial support that it provided me over three years through the Algorithms for Threat Detection program (Grant 1922952) and the URoL (Understanding the Rules of Life: Predicting Phenotype) program (Grant 2124903).

The work in Chapter 3 is a collaboration between Mason Porter and me. Writing with him has been challenging because he expects the best. The result is a paper that I am extremely proud of. Thank you for our many meetings to discuss this work and the best way to present it. I would also like to thank Christopher Klausmeier, Thomas Koffel, Elena Litchman, Sarah Tymochko, and Jonas Wickman for helpful conversations during the early stages of this project and Kedar Karhadkar for helpful discussions about the proofs in this project.

Chapter 5 is adapted from [HJJ24], which I co-authored with Abigail Hickok, Benjamin Jarman, Jiajie Luo, and Mason Porter. This was my first research project at UCLA, and I learned a great deal from all of you. Thank you all for your patience and guidance while I figured out how to conduct research. I would also like to thank Christopher Anderson, Juan Manuel Balcazar, Lyndie Chiou, Alex Sherman, and Renata Turkes for helpful comments and discussions.

On a more personal note, thank you, Ben Jarman, for telling me about careers in quantitative finance. Our discussions took place at a time when I felt uncertain about what life after graduate school would look like. I now have a job lined up at a quantitative finance firm that I am very excited about. Those conversations were the beginning of what has now become the most significant step in my professional life.

I would not be the person I am today if I were not raised to love mathematics and science. I want to thank some incredible teachers and professors who have helped instill this passion over the years. Thank you, Michael Twilling, Tom Vining, Phil Compton, Joshua Pantier, Gary Adams, Matthias Kawski, and Barry Ritchie.

I would now like to thank several people who have been important constants in my personal life. Michael and Delaney Koury, you are the best friends I could ask for. I miss living with you guys, and it's been hard living in different states over the last four years. When graduate school has been exceptionally busy, I have not always been the best at keeping in touch. Despite that, every time I talk to you guys or I'm able to visit in person, I feel as happy and comfortable as I do with anyone in my life. I am worried that starting a career on the East Coast will only make communication more challenging, but I know that you two will always be important parts of my life. I love you and I miss you. Dean, I'm so very excited to meet you.

I have made many new friends at UCLA. For fear of forgetting names, I will not list all of them here, but I appreciate each and every one of you. Going for runs, going to the beach, playing ultimate frisbee, and having Super Bowl parties with you all are memories that I will never forget. I would like to specifically acknowledge Jason Brown, James Chapman, Allison Schiffman, and Yotam Yaniv. You four have been the best part of my time at UCLA, and I love you all very much. Jason, I have missed you since you moved to Seattle. As the rest of us start our careers in different cities, I will miss you a lot, James, Allison, and Yotam. I'm sad that we won't all be living in the same place, but I'm so grateful to have you all as lifelong friends.

I want to thank my family for their love and support throughout my entire life. Jenny, Laura, and Sarah, you are the best sisters in the world. Thanks for always taking the time to see me whenever I come home. I'm excited to see you all again soon. Mom and Dad, thank you for everything. Thank you for raising me to be curious about the world. Thank you for teaching me to be kind. Thank you for your encouragement during good times and

your care during difficult times. Lastly, thank you for supporting my decisions in life, even when they take me to other cities. It's been hard being away. I miss and love you all very much. Tell Skye and Ranger they're the best dogs ever.

Finally, I want to thank my partner Aoife O'Brien. You've been my biggest source of support during my time in graduate school. You inspire me with your kindness and generosity, and you make me want to be the best version of myself. I love you so much. I'm very lucky that I get to spend every day with my best friend, and I'm so excited to start the next chapter of our lives together.

VITA

2016–2020 B.S. in Mathematics. Arizona State University

2016–2020 B.S. in Physics. Arizona State University

2020–2024 M.A. in Mathematics. University of California, Los Angeles

PUBLICATIONS

Abigail Hickok*, Benjamin Jarman*, Michael Johnson*, Jiajie Luo*, and Mason A. Porter. (*joint first author). "Persistent Homology for Resource Coverage: A Case Study of Access to Polling Sites" *SIAM Review*, 66(3):481–500, 2024.

Michael Johnson and Mason A. Porter. "Interacting Hosts with Microbiome Exchange: An Extension of Metacommunity Theory for Discrete Interactions." *In preparation.*

CHAPTER 1

Introduction

In many scientific disciplines, it is important to study networks of interconnected agents [New18]. In this dissertation, we discuss two research projects in network science. In the first project, we introduce a novel modeling framework to study the dynamics of microbiomes of interacting hosts. Using our framework, we investigate the impact of discrete microbiome exchange between hosts on their microbiomes. In the second project, we use a tool from topological data analysis (TDA) called persistent homology (PH) to classify the coverage of resources. We apply our methodology to a case study of polling-site coverage during the 2016 United States presidential election.

1.1 Interacting Hosts with Microbiome Exchange

The microbiomes of humans and other animals play a critical role in their functioning and health [HWC22, VWS18, WYL17]. A classical approach to study such populations is to analyze phenomenological dynamical-systems models [Ede05], such as the generalized Lotka–Volterra model [CIM24]. In these models, one tracks the abundances $N_k(t)$ of different microbe species. These abundances are collected into an abundance vector $\boldsymbol{N}(t)$, and one assumes that the dynamics of each abundance is a function of the abundance vector $\boldsymbol{N}(t)$. Therefore, the dynamics are governed by a dynamical system

$$\frac{\mathrm{d}\boldsymbol{N}}{\mathrm{d}t} = g\left(\boldsymbol{N}\right) . \tag{1.1}$$

These models assume that all microbe species exist in a single environment. Therefore, we refer to these ecological models as models of *local dynamics*.

The specification of local ecological dynamics provides a necessary starting point to study microbiomes, but it is also necessary to account for interactions across different environments, which are essential to understand microbiome composition in many settings [RAT21,SHJ20,TBB15]. Researchers employ *metacommunity theory* [HLH05,LHM04] to investigate the effects of multiple-scale interactions that include both local ecological dynamics and interactions across distinct environments. One relevant framework in metacommunity theory is the *mass-effects paradigm* [LRR23, LMG03, ML02, ML03, TGD20], in which researchers study systems with local ecological dynamics and dispersal between environmental patches. Models in this framework are often coupled differential equations of the form

$$\frac{\mathrm{d}\boldsymbol{N}^{(i)}}{\mathrm{d}t} = g^{(i)}(\boldsymbol{N}^{(i)}) + \sum_j \sigma_{ij} \left(\boldsymbol{N}^{(j)} - \boldsymbol{N}^{(i)}\right) , \tag{1.2}$$

where $\boldsymbol{N}^{(i)}(t)$ is a vector that encodes the microbe species abundances in patch i at time t. The autonomous function $g^{(i)}$ encodes the local dynamics in patch i, and the parameter σ_{ij} governs the dispersal between patches i and j.

Mass-effects models (e.g., see (1.2)) fail to capture an essential aspect of the microbiomes of many living hosts. Many living hosts (such as humans) do not interact continuously [MHJ08] and thus do not sustain a continuous dispersal of microbes. Instead, they interact in discrete time intervals. In Chapter 3, we develop a framework that considers the discrete nature of host interactions. In this framework, when two hosts interact with each other, they instantaneously exchange some of their microbiomes.

We derive analytical approximations of models in our framework in three parameter regimes and prove that they are accurate in those regimes. We also compare these approximations to numerical simulations for an illustrative model. We demonstrate that both parameters in our modeling framework are necessary to determine microbiome dynamics. Key features of the dynamics, such as microbiome convergence across hosts, depend sensi-

tively on the interplay between interaction frequency and strength.

1.2 Persistent Homology for Resource Coverage

The geographical distribution of resources such as polling sites (i.e., locations where people vote), hospitals, COVID-19 vaccination sites, Department of Motor Vehicles (DMV) locations, and Planned Parenthood clinics is a significant factor in the equitability of access to those resources. Consequently, given the locations of a set of resource sites, it is important to quantify their geographical coverage and to identify underserved geographical regions.

One way to study the distribution of these resources is using persistent homology (PH) [OPT17]. PH is built on the theory of *homology* [Hat02], a branch of algebraic topology that characterizes a topological space by its "holes". PH can be used to extend homology to the characterization of *point clouds*. A point cloud is a finite collection $X = \{x_i\}_{i=1}^n$ of points in a metric space (M, d). To study PH, one starts by constructing a *filtered simplicial complex*. A *simplicial complex* is a collection of vertices, edges, triangles, and higher-dimensional simplices. A filtered simplicial complex is a nested sequence $\mathcal{K}_{\alpha_0} \subseteq \mathcal{K}_{\alpha_1} \subseteq \cdots \subseteq \mathcal{K}_{\alpha_n}$ of simplicial complexes, where $\alpha_0 < \alpha_1 < \cdots < \alpha_n$. One example is a Čech filtration [OPT17]. For $r > 0$, the *Čech complex* $\check{\mathrm{C}}_r(X, M, d)$ *at filtration parameter* r is the simplicial complex that has a simplex with vertices $[x_{i_0}, \ldots, x_{i_k}]$ if the intersection $\bigcap_j B(x_{i_j}, r)$ is nonempty, where $B(x, r) := \{y \in M \mid d(x, y) \leq r\}$. By characterizing the homology of a simplicial complex at each filtration level r, one can determine the location and persistence of holes in a point cloud. We discuss background on PH in more detail in Chapter 4.

In Chapter 5, we use PH to identify holes in resource coverage. We construct a distance function d that is based on the travel times to a resource. This distance function is better than geographical distance at encoding the costs of accessing resources. We apply our methodology to study polling-site access for several geographical areas in the 2016 United States presidential election. We analyze and compare the coverage of polling sites in Los

Angeles County and five cities (Atlanta, Chicago, Jacksonville, New York City, and Salt Lake City).

1.3 Organization of the Dissertation

In Chapter 2, we provide background on network-science fundamentals. In Chapter 3, we present our modeling framework for interacting hosts with microbiome exchange. In Chapter 4, we discuss relevant background on PH. In Chapter 5, we develop our PH methodology for the classification of resource coverage and examine the coverage of polling sites in the 2016 US presidential election. Chapter 5 is based on research presented in [HJJ24]. In Chapter 6, we give a few concluding remarks.

CHAPTER 2

Background on Networks

In this chapter, we discuss network-science concepts that are relevant to this dissertation. In Section 2.1, we give various definitions and terminology for networks. In Section 2.2, we discuss differences between synthetic networks and networks that one constructs from real-world data. We also discuss some uses for both synthetic and real-world networks. Finally, in Section 2.3, we briefly describe the networks that we will use in the later chapters of this dissertation.

Throughout the chapter, we frequently cite Mark Newman's textbook *Networks* [New18] and Francesco Bullo's *Lectures on Network Systems* [Bul24] as introductory sources of information.

2.1 Network Fundamentals

The simplest types of networks $G = (V, E)$ consist of a set V of *nodes* and a set $E \subseteq V \times V$ of *edges* between those nodes [Bul24]. These networks are also known as *graphs*; we will use the terms "network" and "graph" interchangeably throughout this dissertation. The nodes of a network can represent agents of some form, and the edges encode ties between the agents. For example, a network can represent users with Facebook accounts as nodes and encode whether or not two users are friends on Facebook in the edge set E. Edges can be either *undirected* or *directed*. In a directed network, each edge has a starting node and an ending node. By convention, we use (i, j) to denote an edge that points from node i to node

j. In an undirected network, each edge is an unordered pair of nodes; there is no notion of a starting node or an ending node. In this case, one represents an edge by writing either (i, j) or (j, i) [New18]. Directed networks are useful when the edges represent asymmetric relationships, such as Instagram following. It is preferable to use an undirected network when the edges represent symmetric relationships, such as Facebook friendships.

In many cases, it is useful to consider edges with *weights*. We refer to a network with edge weights as a *weighted network*. A weighted network is a triplet $G = (V, E, \{w_e\}_{e \in E})$, where w_e is the associated weight of edge e. For an edge $e = (i, j)$, we equivalently write w_e or w_{ij} to represent that edge's associated weight [Bul24]. Weighted edges can encode more information about a relationship between two nodes than unweighted edges. For example, one may wish to have a weight that captures how often people interact on Facebook rather than only tracking whether or not they are friends. Larger weights represent stronger ties between nodes. If, instead, one wants a larger value to indicate a weaker tie between nodes, one uses a *distance*. One can then write $G = (V, E, \{d_e\}_{e \in E})$, where $d_e \in \mathbb{R}_{>0}$ is the associated distance of edge e. For an edge $e = (i, j)$, we write equivalently d_e or d_{ij} to represent that edge's associated distance.

If there is a directed or undirected edge between nodes i and j, we say that these two nodes are *adjacent*. The nodes that are adjacent to i are *neighbors* of i, and the set of such neighbors is the *neighborhood* of i. The number of neighbors of a node i is the degree of i, which we denote by k_i. For a directed network, one also separately tracks the in-degrees and out-degrees. We denote the number of edges that end at node i by k_i^{in}. Analogously, we denote the number of edges that begin at node i by k_i^{out}. A useful way to represent a network is with an *adjacency matrix*. The adjacency matrix A of a network with node set V is a $|V| \times |V|$ matrix. An entry A_{ij} of the adjacency matrix encodes information about the edge (i, j). For an unweighted and undirected network, $A_{ij} = A_{ji} = 1$ if there is an edge between i and j; otherwise, $A_{ij} = A_{ji} = 0$. For a weighted and undirected network, $A_{ij} = A_{ji} = w_{ij}$ if there is an edge between i and j; otherwise, $A_{ij} = A_{ji} = 0$. In a directed network,

we lose symmetry and typically $A_{ij} \neq A_{ji}$. If one uses distances instead of weights, one constructs a *distance matrix* D instead of an adjacency matrix. For an undirected network, $D_{ij} = D_{ji} = d_{ij}$ if there is an edge between i and j; otherwise, $D_{ij} = D_{ji} = \infty$. Analogously to an adjacency matrix, we no longer have symmetry for a directed network and typically $D_{ij} \neq D_{ji}$ [New18].

2.2 Real-World and Synthetic Networks

In many contexts, it is useful to construct a network from real-world data. For example, if you are studying Facebook networks, you may want to use the FACEBOOK100 data set, a collection of Facebook friendship networks of 100 United States universities from fall 2005 [TMP12]. In other contexts, synthetically generated networks are more useful. One reason for this is that collecting real-world data can be impossible or prohibitively difficult. For example, there are several obstacles to constructing a social network that accurately represents the in-person interactions of students at a university [BPP20]. At a large university, the surveying effort that is required to construct such a network costs a considerable amount of both money and time. Even if one is able to survey every student at a university, inaccuracies will still occur in reported interactions. Different students will have different interpretations of what interactions are worth reporting, and many students will likely forget about interactions that they had with other students. Additionally, some pairs of students will report inconsistent frequencies of interaction with each other. A researcher has to make decisions about how to conduct such a survey and how to handle inconsistent responses, and these decisions meaningfully impact the constructed social network. As an alternative, one can also use a proxy, such as a social-media network, for the social network of a university. However, such proxies may not accurately represent social structures that are important to one's research goals.

Another benefit of using synthetically generated networks is that such networks allow

researchers to control the structural properties of networks. For example, in many synthetic network models, one can specify the mean degree of nodes. By varying this parameter, researchers can develop an understanding of how the mean degree affects a phenomenon of interest, such as the qualitative behavior of dynamical processes.

2.3 Networks in this Dissertation

In Chapter 3, we study a model of microbial exchange between hosts. The nodes in our network represent these microbial hosts, and the edges encode which hosts are able to interact with each other. Edge weights encode the frequencies of interactions between pairs of hosts. When two hosts interact, each host exchanges some of its microbiome with the other host. Therefore, we construct the network as an undirected network. We use edge weights to capture the frequency of interactions between hosts.

In Chapter 5, we analyze the coverage of polling sites in six geographical regions during the United States 2016 election. That chapter is adapted from [HJJ24], which was led jointly by Abigail Hickok, Benjamin Jarman, Jiajie Luo, and me and was coauthored with Mason A. Porter. Each network that we construct for this analysis has a city's polling sites as its nodes and an edge between each pair of those nodes. A distance function on the edges encodes the travel times between pairs of nodes. Because travel times can differ based on the direction of travel, the edges in this network are directed.

CHAPTER 3

Interacting Hosts with Microbiome Exchange: An Extention of Metacommunity Theory for Discrete Interactions

In this chapter, we develop a modeling framework for the microbiomes of interacting hosts. This modeling framework is an extension of existing metacommunity frameworks [HLH05, LHM04] that accounts for the discrete nature of host interactions. This chapter is based on in-preparation work that is coauthored with Mason A. Porter.[1]

The chapter proceeds as follows. In Section 3.1, we discuss dynamical-systems models of microbe populations. We also discuss metacommunity theory, which is a set of frameworks that accounts for both the interaction of microbes at a local scale and the exchange of microbes between local environments. These topics provide relevant background necessary for our modeling framework. In Section 3.2, we describe our modeling framework for interacting microbiome hosts. In Section 3.3, we describe the behavior of models in our framework in a regime in which hosts interact with low frequency. In Section 3.4, we describe two distinct regimes in which hosts interact with high frequency. In Section 3.5, we present numerical experiments for our framework. In Section 3.6, we discuss conclusions, limitations, and potential future directions of our work. In Section 3.7, we

[1]I developed the modeling framework and the three approximations that we present in this chapter. I also proved the accuracy of these approximations and wrote the code for the numerical experiments. This work was completed with consistent guidance from and discussions with Mason A. Porter. We are writing the paper together.

prove the accuracy of the approximations in Sections 3.3 and 3.4. Our code is available at `https://github.com/mcjcard/Interacting-Hosts-with-Microbe-Exchange.git`.

3.1 Introduction

The microbiomes of humans and other animals play a critical role in their functioning and health [HWC22, VWS18, WYL17], and there is strong evidence that a host's social interactions significantly impact their microbiome composition [RAT21, SHJ20, TBB15]. Therefore, it is important to study ecological modeling frameworks that account simultaneously for microbe-scale dynamics and the effects of host interactions [AD17, MSB18]. For example, socially determined microbiome signatures are significant indicators of childhood airway development [CHT22], communicable-disease resistance [SMH24], and mental health [KSC22].

3.1.1 Models of Local Ecological Dynamics

The dynamics of microbe populations are affected by environmental factors and the abundances of microbe species. A classical approach to study such populations is to analyze phenomenological dynamical-systems models [Ede05], such as the generalized Lotka–Volterra model [CIM24]

$$\frac{\mathrm{d}N_k}{\mathrm{d}t} = r_k N_k + \sum_{l=1}^{m} \alpha_{kl} N_k N_l \,, \tag{3.1}$$

which describes the dynamics of the abundances $N_1(t), \ldots, N_m(t)$ of m coexisting microbe species. Each species k has an intrinsic birth rate and a death rate, which are combined into a single parameter r_k. A positive r_k signifies that the birth rate exceeds the death rate, and a negative r_k signifies that the death rate exceeds the birth rate. Each cross parameter α_{kl} quantifies the effect of species l on the population of species k. A positive α_{kl} indicates that species l is beneficial to species k, and a negative α_{kl} indicates that species l is harmful to species k.

Researchers also employ mechanistic models, such as consumer–resource models [CIM24], to describe microbial population dynamics. One class of such models is niche models

$$\begin{aligned} \frac{\mathrm{d}N_k}{\mathrm{d}t} &= N_k\, A_k(\boldsymbol{R})\,, \\ \frac{\mathrm{d}R_l}{\mathrm{d}t} &= B_l(\boldsymbol{R}) - \sum_{k=1}^{m} N_k\, C_{kl}(\boldsymbol{R})\,. \end{aligned} \tag{3.2}$$

In a niche model, one tracks both the microbe species' abundances $N_1(t), N_2(t), \ldots, N_m(t)$ and the resource abundances $R_1(t), R_2(t), \ldots, R_n(t)$. The abundance $N_k(t)$ of microbe species k grows or decays according to a growth rate that is a function $A_k(\boldsymbol{R})$ of the resource abundances. The resource abundance $R_l(t)$ is affected both by the resource abundances and by the consumption of the resource by microbes. The function $B_l(\boldsymbol{R})$ encodes the intrinsic dynamics of the resource abundances. The function $C_{kl}(\boldsymbol{R})$ encodes the amount of resource l that species k consumes per unit abundance of species k. Niche models capture the fact that microbes indirectly affect one another through resource competition, rather than interacting directly with each other. Niche models also allow researchers to examine environment-dependent cross-species effects.

3.1.2 Metacommunity Theory

The phenomenological and mechanistic models in Section 3.1.1 assume that all microbe species exist in a single environment. Therefore, we refer to these ecological models as models of *local dynamics*. The specification of local ecological dynamics provides a necessary starting point, but it is also necessary to account for interactions across different environments, which are essential to understand microbiome composition in many settings [RAT21, SHJ20, TBB15]. For example, a coral reef has distinct environmental patches. Stony coral, sponges, algae, and other biotopes provide different conditions to their respective microbiomes. However, the microbiomes of these patches are not isolated and thus impact each other via microbe dispersal [CSC19].

Researchers employ *metacommunity theory* [HLH05, LHM04] to investigate the effects

of multiple-scale interactions that include both local ecological dynamics and interactions across distinct environments. One relevant framework in metacommunity theory is the *mass-effects paradigm* [LRR23, LMG03, ML02, ML03, TGD20], in which researchers study systems with local ecological dynamics and dispersal between environmental patches. Models in this framework are often coupled differential equations of the form

$$\frac{\mathrm{d}\boldsymbol{N}^{(i)}}{\mathrm{d}t} = g^{(i)}(\boldsymbol{N}^{(i)}) + \sum_j \sigma_{ij} \left(\boldsymbol{N}^{(j)} - \boldsymbol{N}^{(i)}\right) , \tag{3.3}$$

where $\boldsymbol{N}^{(i)}(t)$ is a vector that encodes the microbe species abundances in patch i at time t. For consumer–resource models, one can include resources as separate entries of each vector $\boldsymbol{N}^{(i)}(t)$. The autonomous function $g^{(i)}$ encodes the local dynamics in patch i. The parameter σ_{ij} governs the dispersal between patches i and j.

3.1.3 Our Contributions

Mass-effects models (e.g., see (3.3)) fail to capture an essential aspect of the microbiomes of many living hosts. Many living hosts (such as humans) do not interact continuously [MHJ08] and thus do not sustain a continuous dispersal of microbes. Instead, they interact in discrete time intervals. In the present chapter, we develop a framework that considers the discrete nature of host interactions. In this framework, when two hosts interact with each other, they instantaneously exchange some of their microbiomes.

3.2 Our Modeling Framework

In Table 3.1, we summarize the key notation that we use throughout this chapter.

3.2.1 Interaction Network

To study the microbiomes of living hosts, we consider networks that encode the interactions between these hosts. The simplest type of network is a graph $G = (H, E)$, which consists

Table 3.1: Glossary of our Key Notation

Symbols	Definition
H	Node set, which is the set of microbiome hosts (i.e., nodes)
$H^{(i)}$	Microbiome host in H
E	Edge set, which is the set of connections (i.e., edges) between hosts
$\left(H^{(i)}, H^{(j)}\right)$	Edge between $H^{(i)}$ and $H^{(j)}$
λ_{ij}	Interaction-frequency parameter between hosts $H^{(i)}$ and $H^{(j)}$
λ_{tot}	Total-interaction-frequency parameter $\left(\lambda_{\text{tot}} = \sum_{i=1}^{\lvert H\rvert} \sum_{j=i+1}^{\lvert H\rvert} \lambda_{ij}\right)$
l_{ij}	Relative interaction-frequency parameter $\left(l_{ij} = \frac{\lambda_{ij}}{\lambda_{\text{tot}}}\right)$
γ	Interaction strength
$\boldsymbol{N}^{(i)}(t)$	Microbiome abundance vector of host $H^{(i)}$
n	Dimension of each microbiome abundance vector $\boldsymbol{N}^{(i)}(t)$
$\overline{\boldsymbol{N}}$	Mean microbiome abundance vector $\left(\overline{\boldsymbol{N}} = \frac{1}{\lvert H\rvert} \sum_{j=1}^{\lvert H\rvert} \boldsymbol{N}^{(j)}\right)$
$g^{(i)}$	Local-dynamics function of host $H^{(i)}$
$\boldsymbol{\psi}^{(i)}$	Basin probability vector of host $H^{(i)}$
m_i	Dimension of the basin probability vector $\boldsymbol{\psi}^{(i)}$
Ψ	Basin probability tensor
$\Phi^{(ij)}$	Pairwise interaction operator between hosts $H^{(i)}$ and $H^{(j)}$
Φ	total-interaction operator
t^*	Frequency-scaled time ($t^* = \lambda_{\text{tot}} t$)

of a set H of nodes and a set $E \subseteq H \times H$ of edges between nodes. In the context of our modeling framework, we refer to each graph as an *interaction network*. The nodes in the node set H are microbiome hosts. The edges in the edge set $E \subseteq H \times H$ encode whether or not two hosts can interact with each other.

Each edge $e \in E$ also has an associated weight $\lambda_e \in \mathbb{R}^+$. For an edge $e = \left(H^{(i)}, H^{(j)}\right)$, we equivalently write λ_e or λ_{ij}. If there is no edge between hosts $H^{(i)}$ and $H^{(j)}$, we set $\lambda_{ij} = 0$. We refer to this weight as the *interaction-frequency parameter* between hosts $H^{(i)}$ and $H^{(j)}$, as it determines the frequency of the interactions between those two hosts. The order of the hosts in indexing is arbitrary, so $\lambda_{ji} = \lambda_{ij}$. Throughout this chapter, any symbol with the index order ij is equivalent to that symbol with the reverse index order ji. In Figure 3.1, we show an example of an interaction network with ten hosts. We use this network for our numerical experiments in Section 3.5.

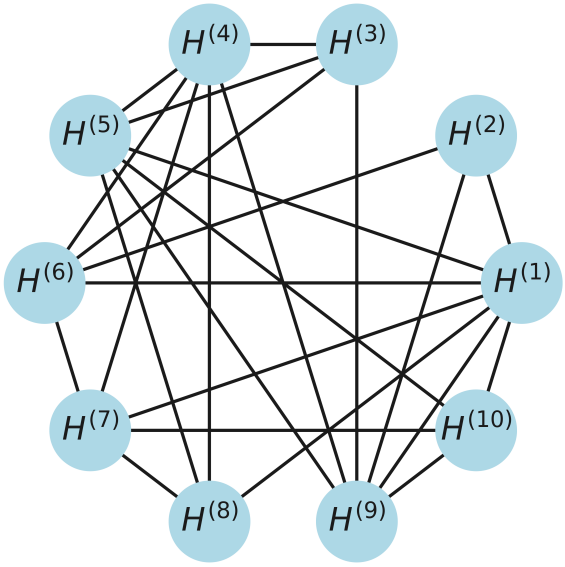


Figure 3.1: An example of an interaction network with 10 hosts. An edge between two hosts indicates that those two hosts can interact with each other. One can represent heterogeneous interaction-frequency parameters λ_{ij} by using different line widths for different edges. In this example, all λ_{ij} values are either 0 or 1.

3.2.2 Exchange Dynamics

Each host $H^{(i)} \in H$ supports a microbiome system. We encode the state of this system with a vector $\boldsymbol{N}^{(i)}(t)$ of microbe species abundances. We refer to $\boldsymbol{N}^{(i)}(t)$ as the *microbiome abundance vector* of host $H^{(i)}$. The kth entry $N_k^{(i)}(t)$ of the microbiome abundance vector encodes the abundance of microbe species k in host $H^{(i)}$ at time t. We order the microbiome abundance vector of each host so that its kth entry describes the same microbe species for each host. The dimension n of all microbiome abundance vectors is the same. If two hosts $H^{(i)}$ and $H^{(j)}$ interact at time t_I, each host instantaneously exchanges a proportion γ of its microbiome with the other host. That is,

$$\begin{aligned} \boldsymbol{N}^{(i)}\left(t_I^+\right) &= (1-\gamma)\boldsymbol{N}^{(i)}\left(t_I^-\right) + \gamma \boldsymbol{N}^{(j)}\left(t_I^-\right) , \\ \boldsymbol{N}^{(j)}\left(t_I^+\right) &= (1-\gamma)\boldsymbol{N}^{(j)}\left(t_I^-\right) + \gamma \boldsymbol{N}^{(i)}\left(t_I^-\right) , \end{aligned} \tag{3.4}$$

where the parameter γ governs the strength of the interaction. At an interaction time t_I, the microbiome abundance vector of each host $H^{(i)}$ satisfies $\boldsymbol{N}^{(i)}\left(t_I\right) = \boldsymbol{N}^{(i)}\left(t_I^+\right)$. For simplicity, we use the same value of the *interaction strength* γ for each pair of hosts. We model the time between consecutive interactions for a pair of adjacent hosts $H^{(i)}$ and $H^{(j)}$ as an exponentially distributed random variable $X_{ij} \sim \text{Exp}(\lambda_{ij})$.

For convenience, we review relevant background information about exponential distributions. For further details, see [Fel68]. The probability density function f_{ij} for an exponential distribution with parameter λ_{ij} is

$$f_{ij}(t) = \lambda_{ij} e^{-\lambda_{ij} t} . \tag{3.5}$$

An exponentially distributed random variable X_{ij} is memoryless. No matter how much time passes after the most recent interaction between hosts $H^{(i)}$ and $H^{(j)}$, the time that remains until the next interaction is distributed as X_{ij}. That is,

$$\Pr(X_{ij} > t + s \mid X_{ij} > s) = \Pr(X_{ij} > t) . \tag{3.6}$$

Because X_{ij} is memoryless, it is easy to describe the random variable

$$X = \min_{i,j>i}\{X_{ij}\}\,, \tag{3.7}$$

for the time until the next interaction between any pair of hosts. The random variable X is also exponentially distributed: $X \sim \mathrm{Exp}(\lambda_{\mathrm{tot}})$, where

$$\lambda_{\mathrm{tot}} = \sum_{i=1}^{|H|} \sum_{j=i+1}^{|H|} \lambda_{ij} \tag{3.8}$$

is the *total-interaction-frequency parameter*. The probability that a given interaction is between a specified pair, $H^{(i)}$ and $H^{(j)}$, of hosts interacts is the *relative interaction-frequency parameter*

$$l_{ij} = \Pr(X_{ij} = X) = \frac{\lambda_{ij}}{\lambda_{\mathrm{tot}}}\,. \tag{3.9}$$

3.2.3 Local Dynamics

Between interactions, an autonomous local dynamical system

$$\frac{\mathrm{d}\boldsymbol{N}^{(i)}}{\mathrm{d}t} = g^{(i)}\left(\boldsymbol{N}^{(i)}\right) \tag{3.10}$$

governs the time evolution of the microbiome abundance vector of each host $H^{(i)}$. This dynamical system encodes the local dynamics of host $H^{(i)}$. We refer to the function $g^{(i)}$ as the *local-dynamics function* of host $H^{(i)}$.

Let the flow $\boldsymbol{X}^{(i)}(t, \boldsymbol{x})$ be the solution of

$$\begin{aligned} \frac{\partial \boldsymbol{X}^{(i)}}{\partial t}(t, \boldsymbol{x}) &= g^{(i)}\left(\boldsymbol{X}^{(i)}(t, \boldsymbol{x})\right)\,, \\ \boldsymbol{X}^{(i)}(0, \boldsymbol{x}) &= \boldsymbol{x}\,. \end{aligned} \tag{3.11}$$

For each $g^{(i)}$, we require that each element of every valid flow is always finite and nonnegative. Specifically, there is a constant $M \in R^+$ such that $\boldsymbol{x} \in [0, M]^n$ implies that each flow $\boldsymbol{X}^{(i)}(t, \boldsymbol{x}) \in [0, M]^n$ for all times $t \geq 0$. When this condition holds, we say each $g^{(i)}$ is *bounded*.

This is a reasonable assumption for microbiome systems because abundances cannot increase without bound or become negative.

When all $g^{(i)}$ are bounded, local dynamics cannot cause any microbiome abundance vector $\boldsymbol{N}^{(i)}(t)$ to leave the region $[0, M]^n$. Interactions also cannot cause any microbiome abundance vector to leave this region. Consider an interaction at time t_I between hosts $H^{(i)}$ and $H^{(j)}$ with microbiome abundance vectors that satisfy $\boldsymbol{N}^{(i)}\left(t_I^-\right) \in [0, M]^n$ and $\boldsymbol{N}^{(j)}\left(t_I^-\right) \in [0, M]^n$. The microbiome exchange between these hosts is an averaging process. After the interaction, the microbiome abundance vectors satisfy $\boldsymbol{N}^{(i)}\left(t_I^+\right) \in [0, M]^n$ and $\boldsymbol{N}^{(j)}\left(t_I^+\right) \in [0, M]^n$. Therefore, if each $g^{(i)}$ is bounded and each $\boldsymbol{N}^{(i)}(0) \in [0, M]^n$, it follows that each $\boldsymbol{N}^{(i)}(t) \in [0, M]^n$ for all times $t \geq 0$.

We now present an illustrative model of local dynamics that we use repeatedly to illustrate our framework. For this illustrative model, we assume that each host sustains two microbe species. Therefore, each microbiome abundance vector has dimension two. We use the dynamical system

$$\begin{aligned} \frac{\mathrm{d}N_1^{(i)}}{\mathrm{d}t} &= -\frac{N_1^{(i)}}{10}\left(N_1^{(i)} - 2\right)\left(N_1^{(i)} - 8\right)\left(N_1^{(i)} - 12\right), \\ \frac{\mathrm{d}N_2^{(i)}}{\mathrm{d}t} &= -\frac{N_2^{(i)}}{10}\left(N_2^{(i)} - 2\right)\left(N_2^{(i)} - 11\right)\left(N_2^{(i)} - 12\right) \end{aligned} \tag{3.12}$$

for the local dynamics of each host. In Figure 3.2, we show this dynamical system's four stable equilibrium points and their associated basins of attraction. We use the labels 1, 2, 3, and 4 for the basins of attraction of the attractors $(2, 2)$, $(12, 2)$, $(2, 12)$, and $(12, 12)$, respectively.

Consider an interaction network with two hosts $H^{(1)}$ and $H^{(2)}$ that are connected by a single edge. Suppose that there is an interaction between the two hosts at time t_I and that $\boldsymbol{N}^{(1)}(t_I^-) = (2, 2)$ and $\boldsymbol{N}^{(2)}(t_I^-) = (12, 12)$. In Figure 3.3, we show the states of the two hosts immediately after interacting for three values of the interaction strength γ. This figure demonstrates that a single interaction can change the basin of attraction of a host's microbiome abundance vector for sufficiently large γ.

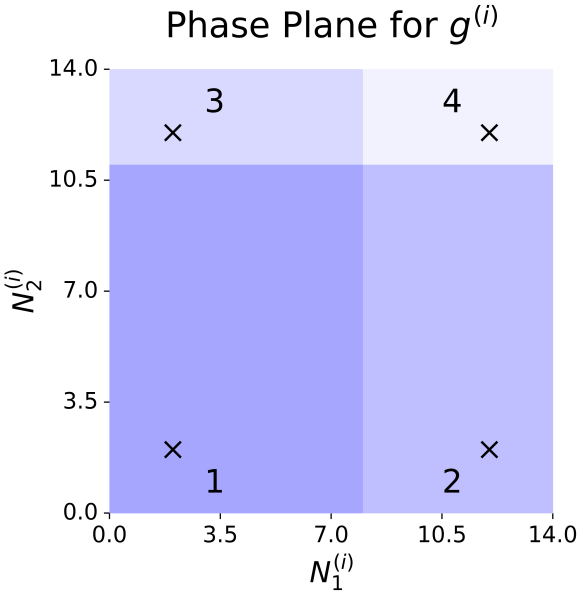


Figure 3.2: The four stable equilibrium points for the illustrative model (3.12) of local dynamics and their basins of attraction. We use the labels 1, 2, 3, and 4 for the basins of attraction of the attractors $(2, 2)$, $(12, 2)$, $(2, 12)$, and $(12, 12)$, respectively.

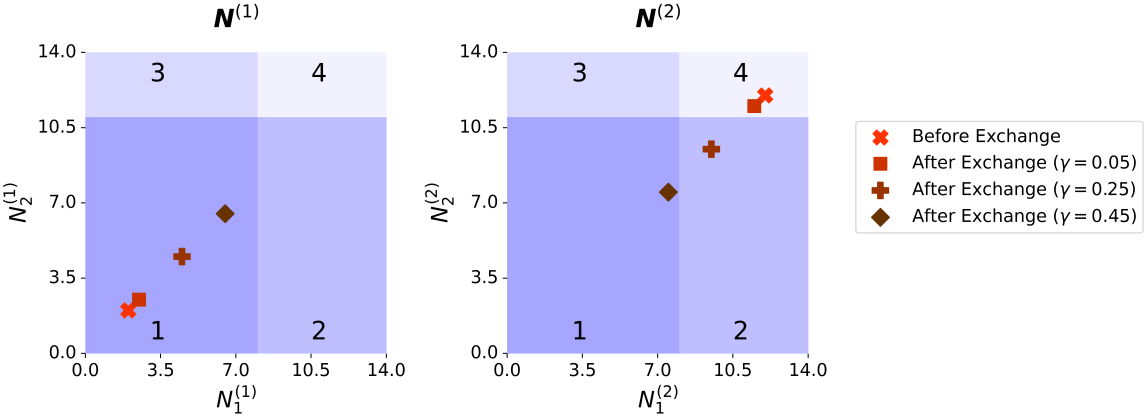


Figure 3.3: Two hosts with local dynamics (3.12). Immediately before interacting at time t_I, the hosts have microbiome abundance vectors $\boldsymbol{N}^{(1)}\left(t_I^-\right) = (2, 2)$ and $\boldsymbol{N}^{(2)}\left(t_I^-\right) = (12, 12)$. We show the microbiome abundance vectors $\boldsymbol{N}^{(1)}\left(t_I^+\right)$ and $\boldsymbol{N}^{(2)}\left(t_I^+\right)$ of the two hosts immediately after interacting for interaction strengths $\gamma = 0.05$, $\gamma = 0.25$, and $\gamma = 0.45$.

If interactions occur in sufficiently quick succession, then smaller values of γ can also cause transitions in the basin of attraction of a host's microbiome abundance vector. In Figure 3.4, we show an example of this phenomenon. We begin with microbiome abundance vectors $\boldsymbol{N}^{(1)}(0) = (2, 2)$ and $\boldsymbol{N}^{(2)}(0) = (12, 12)$, and we track the microbiome abundance vectors through five interactions between the hosts. These interactions occur at times 0.1, 0.3, 0.4, 0.7, and 0.73. In this example, the interaction strength is $\gamma = 0.32$. The first interaction is sufficient to move $\boldsymbol{N}^{(2)}(t)$ from basin 4 to basin 2. However, for $\boldsymbol{N}^{(2)}(t)$ to move from basin 2 to basin 1, two interactions must occur in sufficiently quick succession. In Figure 3.4, we see that interactions that occur at times 0.3 and 0.4 do not cause this transition. However, interactions at times 0.7 and 0.73 are close enough in time to cause $\boldsymbol{N}^{(2)}(t)$ to move from basin 2 to basin 1.

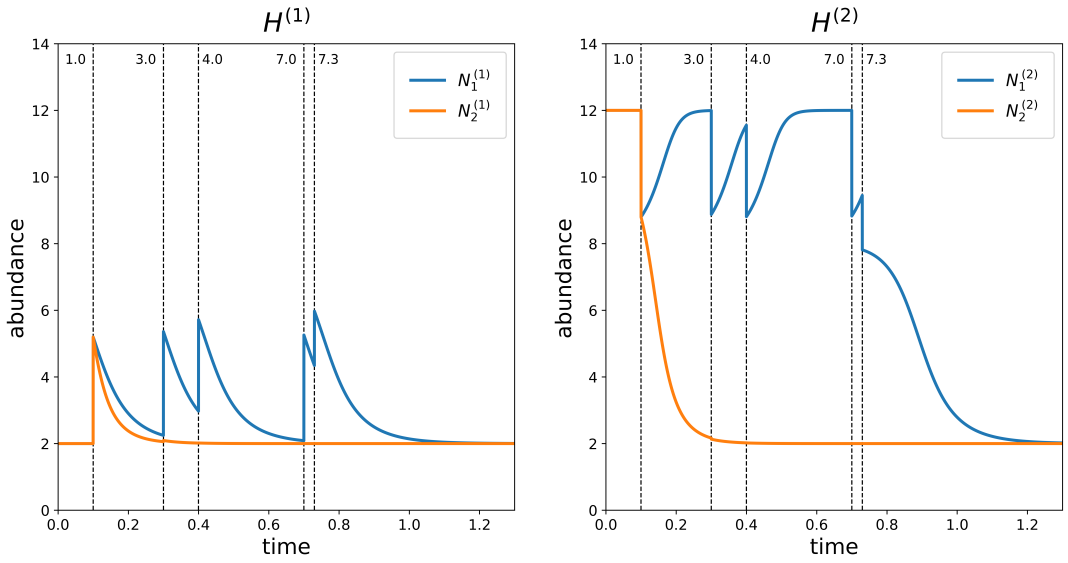


Figure 3.4: Two hosts with local dynamics (3.12). These hosts have initial states $\boldsymbol{N}^{(1)}(0) = (2, 2)$ and $\boldsymbol{N}^{(2)}(0) = (12, 12)$. We show the abundances of microbe species 1 and 2 in each host through the course of five interactions at times 0.1, 0.3, 0.4, 0.7, and 0.73.

3.3 Low-Frequency Approximation (LFA)

In this section, we discuss an approximation that is accurate when all interaction-frequency parameters λ_{ij} are sufficiently small. We develop this *low-frequency approximation* (LFA) for systems in which the set of stable attractors of each host's local dynamics is a finite set of equilibrium points. We believe that it is possible to derive extensions of the LFA for systems with other types of attractors, and we discuss this possibility in Section 3.6.2. We define relevant terminology in Sections 3.3.1 and 3.3.2, and we describe the LFA and outline the proof of its accuracy in Section 3.3.3. We give a complete proof in Section 3.7.1.

3.3.1 Basin State Tensor

We illustrated in Figure 3.3 that interactions can result in transitions of the basin of attraction of a host's microbiome abundance vector. Local dynamics cannot cause such a transition to occur, so interactions between hosts are necessary for such transitions.

Throughout the rest of Section 3.3, we need to be able to track the basin of attraction of a host's microbiome abundance vector. To do this, we define a *basin probability vector* $\boldsymbol{\psi}^{(i)}(t)$ for each host $H^{(i)}$. An entry $\psi_a^{(i)}(t)$ of this vector gives the probability that the microbiome abundance vector of host $H^{(i)}$ is in basin of attraction a at time t. If the local dynamics of host $H^{(i)}$ has m_i basins of attraction, then $\boldsymbol{\psi}^{(i)}(t)$ has dimension m_i. Different hosts can have different local dynamics, so the basin probability vectors of different hosts can have different dimensions.

Each local-dynamics function $g^{(i)}$ is bounded, as described in Section 3.2.3. Therefore, for all times $t \geq 0$, each microbiome abundance vector $\boldsymbol{N}^{(i)}(t) \in [0, M]^n$. Because the set of stable attractors of each host's local dynamics consists of a finite set of equilibrium points, the set of points $\mathcal{U}^{(i)} \subset [0, M]^n$ that are not in the basin of attraction of some equilibrium point has measure 0. For the LFA to be accurate, we require specific conditions on the local-dynamics functions $g^{(i)}$ and the interaction strength γ. We describe these conditions in

the Low-Frequency Approximation Theorem (see Theorem 1). When these conditions on $g^{(i)}$ and γ are satisfied and the total-interaction-frequency parameter $\lambda_{\text{tot}} \to 0$, no microbiome abundance vector $\boldsymbol{N}^{(i)}(t)$ lies on the border between basins of attraction at any time t in a finite interval $[0, T]$ with arbitrarily high probability.

We represent the state of the entire set of hosts using a *basin probability tensor* $\Psi(t)$. If each $\boldsymbol{\psi}^{(i)}(t)$ has dimension m_i, then $\Psi(t)$ has dimension $m_1 \times m_2 \times \cdots \times m_{|H|}$. An entry of the basin probability tensor $\Psi_{a_1,a_2,\ldots,a_{|H|}}(t)$ gives the probability that the microbiome abundance vector of each host $H^{(i)}$ is in basin of attraction a_i at time t. If we assume that these probabilities are independent at time t, then

$$\Psi(t) = \bigotimes_i \boldsymbol{\psi}^{(i)}(t)\,. \tag{3.13}$$

Typically, the $\boldsymbol{\psi}^{(i)}(t)$ are not independent after any interaction, and then (3.13) no longer holds.

3.3.2 Interaction Operators

The basin probability tensor $\Psi(t)$ can change only due to an interaction. Unfortunately, knowing only $\Psi(t)$ before an interaction and which pair of hosts interacted is insufficient to determine $\Psi(t)$ after the interaction. One also needs information about the microbiome abundance vectors of the interacting hosts.

Suppose that an interaction occurs between hosts $H^{(1)}$ and $H^{(2)}$ at time t_I and that their microbiome abundance vectors $\boldsymbol{N}^{(1)}\left(t_I^-\right)$ and $\boldsymbol{N}^{(2)}\left(t_I^-\right)$ are at stable equilibrium points immediately before the interaction. We make this assumption throughout the rest of this section. If we know that the basins of attraction of $\boldsymbol{N}^{(1)}\left(t_I^-\right)$ and $\boldsymbol{N}^{(2)}\left(t_I^-\right)$ are a_1 and a_2, respectively, then we are able to determine the basins of attraction of $\boldsymbol{N}^{(1)}\left(t_I^+\right)$ and $\boldsymbol{N}^{(2)}\left(t_I^+\right)$. We illustrate this in Figure 3.5 for an example in which both hosts have local dynamics (3.12) and the interaction strength is $\gamma = 0.25$. We show $\boldsymbol{N}^{(1)}\left(t_I^+\right)$ for every possible combination of $\boldsymbol{N}^{(1)}\left(t_I^-\right)$ and $\boldsymbol{N}^{(2)}\left(t_I^-\right)$. Because $\boldsymbol{N}^{(1)}\left(t_I^-\right)$ and $\boldsymbol{N}^{(2)}\left(t_I^-\right)$ are at

stable equilibrium points (by assumption), there are 16 such combinations. For example, for $\boldsymbol{N}^{(1)}\left(t_I^-\right) = (12, 12)$, there are four possible values of $\boldsymbol{N}^{(1)}\left(t_I^+\right)$. The values are $(9.5, 9.5)$, $(12, 9.5)$, $(9.5, 12)$, and $(12, 12)$; there are four corresponding values ($(2, 2)$, $(12, 2)$, $(2, 12)$, and $(12, 12)$) of $\boldsymbol{N}^{(2)}\left(t_I^-\right)$. In Figure 3.5, we mark these four possible values of $\boldsymbol{N}^{(1)}\left(t_I^+\right)$ with the diamonds in the upper-right corner. Because host $H^{(2)}$ has the same local dynamics as host $H^{(1)}$, the situation is identical for $\boldsymbol{N}^{(2)}\left(t_I^+\right)$, except that we exchange the indices 1 and 2 everywhere.

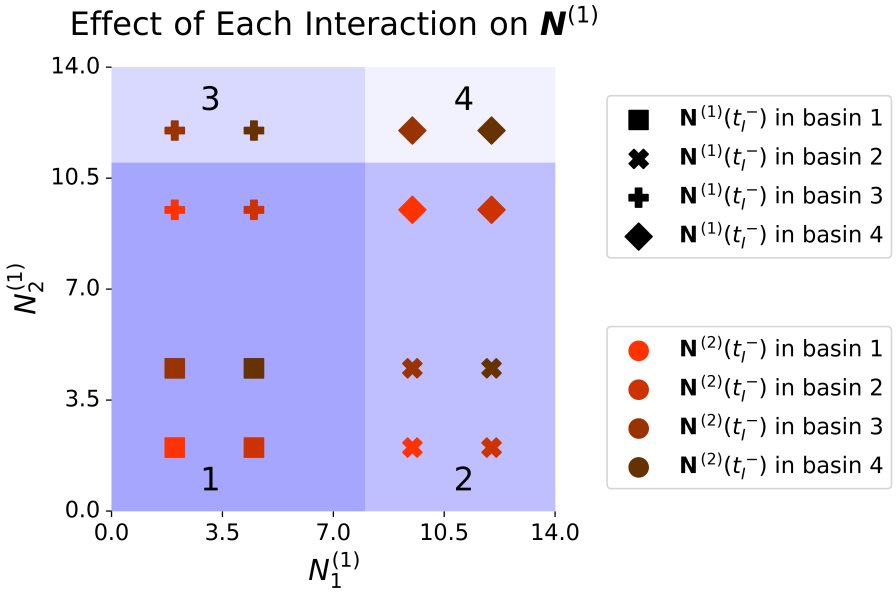


Figure 3.5: Each possible microbiome abundance vector $\boldsymbol{N}^{(1)}\left(t_I^+\right)$ after an interaction at time t_I between hosts $H^{(1)}$ and $H^{(2)}$ with local dynamics (3.12), assuming that $\boldsymbol{N}^{(1)}\left(t_I^-\right)$ and $\boldsymbol{N}^{(2)}\left(t_I^-\right)$ are at stable equilibrium points before the interaction. The marker shapes indicate the basins of attraction of $\boldsymbol{N}^{(1)}\left(t_I^-\right)$, and the marker colors indicate the basins of attraction of $\boldsymbol{N}^{(2)}\left(t_I^-\right)$. For example, for $\boldsymbol{N}^{(1)}\left(t_I^-\right) = (12, 12)$, there are four possible values of $\boldsymbol{N}^{(1)}\left(t_I^+\right)$. These values are $(9.5, 9.5)$, $(12, 9.5)$, $(9.5, 12)$, and $(12, 12)$; there are four corresponding values ($(2, 2)$, $(12, 2)$, $(2, 12)$, and $(12, 12)$) of $\boldsymbol{N}^{(2)}\left(t_I^-\right)$. We mark these four possible values of $\boldsymbol{N}^{(1)}\left(t_I^+\right)$ with the diamonds in the upper-right corner. The color indicates the corresponding value of $\boldsymbol{N}^{(2)}\left(t_I^-\right)$.

For some values of the interaction strength γ, an interaction between hosts $H^{(i)}$ and $H^{(j)}$ can result in either $\boldsymbol{N}^{(i)}\left(t_I^+\right)$ or $\boldsymbol{N}^{(j)}\left(t_I^+\right)$ lying on a boundary between basins of attraction, rather than inside a basin of attraction. That is, after such an interaction, we have $\boldsymbol{N}^{(i)}\left(t_I^+\right) \in \mathcal{U}^{(i)}$ or $\boldsymbol{N}^{(j)}\left(t_I^+\right) \in \mathcal{U}^{(j)}$. We refer to the set of such interaction strengths as the *boundary set* $\mathcal{B}_{ij}$ for the hosts $H^{(i)}$ and $H^{(j)}$. For the example in Figure 3.5, the boundary set is $\mathcal{B}_{12} = \{0.1, 0.4\}$. If hosts $H^{(i)}$ and $H^{(j)}$ cannot interact, then $\mathcal{B}_{ij}$ is the empty set. The *total boundary set* is

$$\mathcal{B} = \bigcup_{i,j} \mathcal{B}_{ij} \,. \tag{3.14}$$

For the LFA, we require $\gamma \notin \mathcal{B}$, and we assume that this is the case for the rest of this section.

Under our assumptions, we need to know only the basins of attraction of $\boldsymbol{N}^{(i)}\left(t_I^-\right)$ and $\boldsymbol{N}^{(j)}\left(t_I^-\right)$ (i.e., before an interaction) to determine the basins of attraction of $\boldsymbol{N}^{(i)}\left(t_I^+\right)$ and $\boldsymbol{N}^{(j)}\left(t_I^+\right)$ (i.e., after the interaction). Therefore, we need to know only the basin probability tensor $\Psi\left(t_I^-\right)$ prior to an interaction to determine the basin probability tensor $\Psi\left(t_I^+\right)$ after that interaction. To describe such an interaction-induced change, we define a *pairwise interaction operator* $\Phi^{(ij)}$ with dimension $m_1 \times \cdots \times m_{|H|} \times m_1 \times \cdots \times m_{|H|}$. Its entry $\Phi^{(ij)}_{b_1,\ldots,b_{|H|},a_1,\ldots,a_{|H|}} = 1$ if b_i and b_j are the basins of attraction of $\boldsymbol{N}^{(i)}\left(t_I^+\right)$ and $\boldsymbol{N}^{(j)}\left(t_I^+\right)$, respectively, when a_i and a_j are the respective basins of attraction of $\boldsymbol{N}^{(i)}\left(t_I^-\right)$ and $\boldsymbol{N}^{(j)}\left(t_I^-\right)$ and $b_k = a_k$ for all $k \notin \{i, j\}$. Otherwise, $\Phi^{(ij)}_{b_1,\ldots,b_{|H|},a_1,\ldots,a_{|H|}} = 0$.

Consider a two-host interaction network in which both hosts have local dynamics (3.12) and the interaction strength is $\gamma = 0.25$. The basin probability tensor of this system has dimension 4×4. Therefore, the pairwise interaction operator $\Phi^{(12)}$ has dimension $4\times4\times4\times4$. In Figure 3.5, we show the results of all possible interactions between these two hosts. For example, if $\boldsymbol{N}^{(1)}\left(t_I^-\right)$ and $\boldsymbol{N}^{(2)}\left(t_I^-\right)$ are in basins of attraction 1 and 4, respectively, then the resulting basins of attraction of $\boldsymbol{N}^{(1)}\left(t_I^+\right)$ and $\boldsymbol{N}^{(2)}\left(t_I^+\right)$ are 1 and 2, respectively. Therefore, $\Phi^{(12)}_{1,2,1,4} = 1$. Because there are 16 possible combinations of $\boldsymbol{N}^{(1)}\left(t_I^-\right)$ and $\boldsymbol{N}^{(2)}\left(t_I^-\right)$, there

are 16 entries of $\Phi^{(12)}$ that equal 1. We list these 16 entries in Figure 3.6.

$$\begin{array}{cccc}
\Phi^{(12)}_{1,1,1,1} & \Phi^{(12)}_{1,2,1,2} & \Phi^{(12)}_{1,1,1,3} & \Phi^{(12)}_{1,2,1,4} \\
\Phi^{(12)}_{2,1,2,1} & \Phi^{(12)}_{2,2,2,2} & \Phi^{(12)}_{2,1,2,3} & \Phi^{(12)}_{2,2,2,4} \\
\Phi^{(12)}_{1,1,3,1} & \Phi^{(12)}_{1,2,3,2} & \Phi^{(12)}_{3,3,3,3} & \Phi^{(12)}_{3,4,3,4} \\
\Phi^{(12)}_{2,1,4,1} & \Phi^{(12)}_{2,2,4,2} & \Phi^{(12)}_{4,3,4,3} & \Phi^{(12)}_{4,4,4,4}
\end{array}$$

Figure 3.6: The 16 entries of the pairwise interaction operator that equal 1 for a two-host interaction network in which both hosts have local dynamics (3.12) and the interaction strength is $\gamma = 0.25$.

We use the Einstein summation convention [SW] to describe how $\Phi^{(ij)}$ operates on the basin probability tensor. In this convention, one sums over any repeated index that occurs in a single term. For example, we describe the product

$$\boldsymbol{y} = A\boldsymbol{x} \tag{3.15}$$

of a matrix A and a vector $\boldsymbol{x}$ by writing

$$y_i = \sum_j A_{ij} x_j \,. \tag{3.16}$$

Using the Einstein summation convention, we write

$$y_i = A_{ij} x_j \,. \tag{3.17}$$

We write the effect of the pairwise interaction operator $\Phi^{(ij)}$ on the basin probability tensor Ψ as

$$\Psi_{b_1,\ldots,b_{|H|}}\left(t_I^+\right) = \Phi^{(ij)}_{b_1,\ldots,b_{|H|},a_1,\ldots,a_{|H|}} \Psi_{a_1,\ldots,a_{|H|}}\left(t_I^-\right) . \tag{3.18}$$

Each $\Phi^{(ij)}$ is a linear operator on the basin probability tensor. At any time, the next interaction is between hosts $H^{(i)}$ and $H^{(j)}$ with probability l_{ij} (see (3.9)). This yields the *total-interaction operator*

$$\Phi = \sum_{i=1}^{|H|} \sum_{j=i+1}^{|H|} l_{ij} \Phi^{(ij)} . \tag{3.19}$$

If we now assume that an interaction occurs between some pair of hosts at time t_I (with probabilities l_{ij} for each pair) and that all $\boldsymbol{N}^{(i)}\left(t_I^-\right)$ are at a stable equilibrium point, then we obtain

$$\Psi_{b_1,\ldots,b_{|H|}}\left(t_I^+\right) = \Phi_{b_1,\ldots,b_{|H|},a_1,\ldots,a_{|H|}} \Psi_{a_1,\ldots,a_{|H|}}\left(t_I^-\right) . \tag{3.20}$$

3.3.3 Low-Frequency-Approximation Theorem

The LFA encodes the evolution of the system (3.4, 3.10) when the local dynamics of each host is much faster than the exchange dynamics between hosts. In this regime, each microbiome abundance vector $\boldsymbol{N}^{(i)}(t)$ becomes close to a stable equilibrium point before the next interaction that involves host $H^{(i)}$. Therefore, the total-interaction operator Φ accurately describes the dynamics of the basin probability tensor $\Psi(t)$.

Before we state the LFA Theorem, we introduce some helpful terminology. The expected number of host interactions in a time interval of duration Δt is $\lambda_{\text{tot}}\Delta t$. We refer to $t^* = \lambda_{\text{tot}} t$ as the *frequency-scaled time*. We say that the local-dynamics function $g^{(i)}$ is *inward pointing* at a point $\boldsymbol{x}$ if there exists a constant $\delta > 0$ such that $\|\boldsymbol{y} - \boldsymbol{x}\|_2 \leq \delta$ implies that $g^{(i)}(\boldsymbol{y}) \cdot (\boldsymbol{x} - \boldsymbol{y}) > 0$.

Theorem 1 (Low-Frequency-Approximation Theorem)**.** *Suppose that the attractors of each host's local dynamics consist of a finite set of stable equilibrium points at which the local-*

dynamics function $g^{(i)}$ is inward pointing, and let each $g^{(i)}$ be continuous and bounded (see Section 3.2.3). Fix $\gamma \notin \mathcal{B}$, all l_{ij}, and a frequency-scaled time T^*. As $\lambda_{\text{tot}} \to 0$, the basin probability tensor $\Psi(t^*)$ converges uniformly to $\widetilde{\Psi}(t^*)$ on $[0, T^*]$, where

$$\frac{d}{dt^*}\widetilde{\Psi}_{b_1,\ldots,b_{|H|}}(t^*) = \Phi_{b_1,\ldots,b_{|H|},a_1,\ldots,a_{|H|}}\widetilde{\Psi}_{a_1,\ldots,a_{|H|}}(t^*) - \widetilde{\Psi}_{b_1,\ldots,b_{|H|}}(t^*)\,, \tag{3.21}$$

$$\widetilde{\Psi}(0) = \Psi(0)\,.$$

We provide key steps of the proof of Theorem 1 in this section. We give a full proof in Section 3.7.1.

As we described in Section 3.3.2, if each $\boldsymbol{N}^{(i)}\left(t_I^-\right)$ is at a stable equilibrium point, then the interaction operator Φ describes the update of the basin probability tensor after an interaction at time t_I. We can construct neighborhoods around each stable equilibrium point of each host's local dynamics such that if each $\boldsymbol{N}^{(i)}\left(t_I^-\right)$ is in one of these neighborhoods, then Φ perfectly describes the transition of $\Psi(t)$ due to an interaction at time t_I. In Figure 3.7, we show an example of what these neighborhoods can look like around the stable equilibrium points for our illustrative model (3.12) of local dynamics.

After an interaction, there is an upper bound on the time that it takes for each microbiome abundance vector to re-enter a neighborhood around a stable equilibrium point. Because the system is finite, there is an upper bound τ such that if no interactions occur in the time interval $[t, t+\tau]$, each microbiome abundance vector $\boldsymbol{N}^{(i)}(t+\tau)$ is in one of these neighborhoods. If no two interactions occur within time τ of each other, then successive applications of the total-interaction operator Φ perfectly describe the evolution of the basin probability tensor $\Psi(t)$. In the frequency-scaled time interval $[0, T^*]$, the expected number of system interactions is T^*. As $\lambda_{\text{tot}} \to 0$, the frequency-scaled time $\tau^* \to 0$. Therefore, it becomes vanishingly unlikely that any pair of interactions occurs within a frequency-scaled time that is less than τ^*. Consequently, as $\lambda_{\text{tot}} \to 0$, the effect on the basin probability tensor $\Psi(t^*)$ of all interactions in $[0, T^*]$ is described perfectly by the total-interaction operator Φ with arbitrarily high probability. Because the interactions are exponentially distributed, the

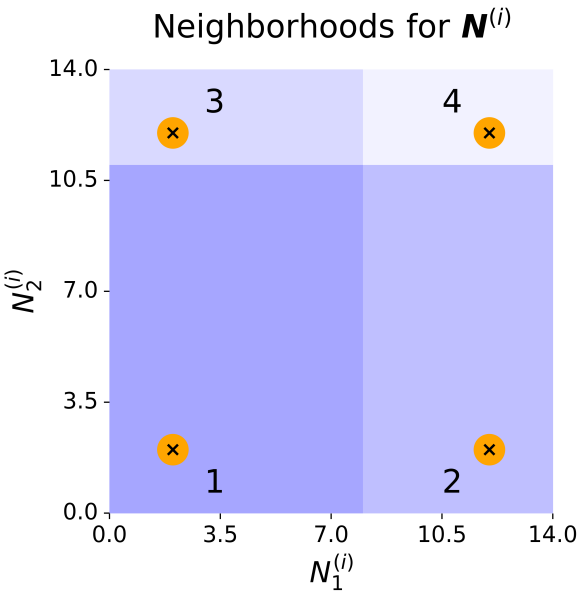


Figure 3.7: An illustration of potential neighborhoods around the four stable equilibrium points for a host with the local dynamics (3.12). These neighborhoods illustrate potential neighborhoods from Theorem 1; they are not the neighborhoods for any particular value of the interaction strength γ.

basin probability tensor $\Psi(t^*)$ converges uniformly to $\widetilde{\Psi}(t^*)$ (see (3.21)).

3.4 High-Frequency Approximations

In this section, we discuss two approximations that are accurate for different regimes with large λ_{tot}. The first of these approximations is the *high-frequency, low-strength approximation* (HFLSA). The HFLSA becomes increasingly accurate as $\lambda_{\text{tot}} \to \infty$ and $\gamma \to 0$ for fixed relative interaction-frequency parameters l_{ij} and fixed $\lambda_{\text{tot}}\gamma$. This approximation results in a model that has the same form as the mass-effects model (3.3). The second approximation is the *high-frequency, constant-strength approximation* (HFCSA). This approximation becomes increasingly accurate as $\lambda_{\text{tot}} \to \infty$ for fixed relative interaction-frequency parameters l_{ij} and fixed interaction strength γ.

3.4.1 High-Frequency, Low-Strength Approximation (HFLSA)

The HFLSA is accurate when the interactions between hosts are very frequent but very weak. In this regime, the expectation of the exchange dynamics (3.4) is constant, but the variance of the exchange dynamics is small. This results in an approximate model for the dynamics of each microbiome abundance vector $\boldsymbol{N}^{(i)}(t)$ that is deterministic and has terms that encode the effects of the local dynamics and the exchange dynamics. This approximate model has the same form as the mass-effects model (3.3).

Theorem 2 (High-Frequency, Low-Strength Approximation Theorem)**.** *Fix the relative interaction-frequency parameters l_{ij}, the product $\lambda_{\text{tot}}\gamma$, and a time T. Let each local-dynamics function $g^{(i)}$ be continuously differentiable and bounded (see Section 3.2.3), and let $\varepsilon \in (0, 1]$ and $\delta > 0$ be arbitrary but fixed. For sufficiently large λ_{tot}, each host's microbiome abundance vector $\boldsymbol{N}^{(i)}(t)$ satisfies*

$$\left\| \boldsymbol{N}^{(i)} - \widetilde{\boldsymbol{N}}^{(i)} \right\|_{L^\infty[0,T]} < \delta \tag{3.22}$$

with probability larger than $1-\varepsilon$*, where*

$$\frac{d\widetilde{\boldsymbol{N}}^{(i)}}{dt} = g^{(i)}(\widetilde{\boldsymbol{N}}^{(i)}) + \sum_j \lambda_{ij}\gamma\left(\widetilde{\boldsymbol{N}}^{(j)} - \widetilde{\boldsymbol{N}}^{(i)}\right), \tag{3.23}$$

$$\widetilde{\boldsymbol{N}}^{(i)}(0) = \boldsymbol{N}^{(i)}(0).$$

We provide key steps of the proof of Theorem 2 in this section. We give a full proof in Section 3.7.2.

Consider the evolution of a microbiome abundance vector $\boldsymbol{N}^{(i)}(t)$ over a short time interval $[t', t'+dt]$. Without interactions, the effect of the local dynamics over this interval is

$$\left(\boldsymbol{N}^{(i)}(t'+dt) - \boldsymbol{N}^{(i)}(t')\right)_{\text{loc}} = g^{(i)}\left(\boldsymbol{N}^{(i)}(t')\right)dt + O(dt^2). \tag{3.24}$$

Let

$$J^{(i)}(t_I) = \boldsymbol{N}^{(i)}\left(t_I^+\right) - \boldsymbol{N}^{(i)}\left(t_I^-\right) \tag{3.25}$$

be the effect of an interaction at time t_I on $\boldsymbol{N}^{(i)}(t)$. If the interaction does not involve $H^{(i)}$, then $J^{(i)}(t_I) = \mathbf{0}$. Otherwise, if the interaction is between $H^{(i)}$ and $H^{(j)}$, then

$$J^{(i)}(t_I) = \gamma\left(\boldsymbol{N}^{(j)}\left(t_I^-\right) - \boldsymbol{N}^{(i)}\left(t_I^-\right)\right). \tag{3.26}$$

Suppose that no interactions occur precisely at times t' or $t'+dt$ and that L interactions occur during the interval $(t', t'+dt)$. We denote this ordered set of interactions by $\{t_l\}_{l=1}^{L}$. Suppose for all i that the microbiome abundance vector $\boldsymbol{N}^{(i)}(t)$ changes very little over the interval $[t', t'+dt]$ such that each $J^{(i)}(t_l)$ is well-approximated by

$$\widetilde{J}_l^{(i)} = \begin{cases} \mathbf{0} & \text{if the interaction at } t_I \text{ does not involve host } H^{(i)} \\ \gamma\left(\boldsymbol{N}^{(j)}(t') - \boldsymbol{N}^{(i)}(t')\right) & \text{if the interaction at } t_I \text{ is between hosts } H^{(i)} \text{ and } H^{(j)}. \end{cases} \tag{3.27}$$

It then follows that the effect of the interactions on the microbiome abundance vector $\boldsymbol{N}^{(i)}(t)$ during this interval is well-approximated by

$$\left(\boldsymbol{N}^{(i)}(t'+dt) - \boldsymbol{N}^{(i)}(t')\right)_{\text{exch}} = \sum_{l=1}^{L} \widetilde{J}_l^{(i)}, \tag{3.28}$$

which we henceforth call the *approximate interaction effect*.

Each of the approximate interaction effects $\widetilde{J}_l^{(i)}$ is vector-valued. We denote entry x of this vector by $\left(\widetilde{J}_l^{(i)}\right)_x$. The sum (3.28) is also vector-valued, and we denote entry x of this sum by $\left(\sum_{l=1}^{L} \widetilde{J}_l^{(i)}\right)_x$. We now calculate the expectation and the variance of each entry of (3.28). The stochasticity in (3.28) arises both from the number L of interactions and from which pair of hosts interact at each time. The number of interactions follows a Poisson distribution with mean $\lambda_{\text{tot}} dt$. The approximate interaction effects $\widetilde{J}_l^{(i)}$ are independent of one another. An interaction at time t_l is between hosts $H^{(i)}$ and $H^{(j)}$ with probability $\lambda_{ij}/\lambda_{\text{tot}}$. Therefore,

$$\begin{aligned}
\mathbb{E}\left[L\right] &= \lambda_{\text{tot}} dt\,, \\
\text{Var}\left[L\right] &= \lambda_{\text{tot}} dt\,, \\
\mathbb{E}\left[\left(\widetilde{J}_l^{(i)}\right)_x\right] &= \sum_j \frac{\lambda_{ij}}{\lambda_{\text{tot}}} \gamma \left(\boldsymbol{N}^{(j)}(t') - \boldsymbol{N}^{(i)}(t')\right)_x\,, \\
\text{Var}\left[\left(\widetilde{J}_l^{(i)}\right)_x\right] &= \mathbb{E}\left[\left(\widetilde{J}_l^{(i)}\right)_x^2\right] - \left(\mathbb{E}\left[\left(\widetilde{J}_l^{(i)}\right)_x\right]\right)^2 .
\end{aligned} \tag{3.29}$$

The expectation of each entry of the sum is

$$\begin{aligned}
\mathbb{E}\left[\left(\sum_{l=1}^{L} \widetilde{J}_l^{(i)}\right)_x\right] &= \mathbb{E}\left[\ \mathbb{E}\left[\left(\sum_{l=1}^{L} \widetilde{J}_l^{(i)}\right)_x \,\middle|\, L\right]\ \right] \\
&= \mathbb{E}\left[L \sum_j \frac{\lambda_{ij}}{\lambda_{\text{tot}}} \gamma \left(\boldsymbol{N}^{(j)}(t') - \boldsymbol{N}^{(i)}(t')\right)_x\right] \\
&= \lambda_{\text{tot}} dt \sum_j \frac{\lambda_{ij}}{\lambda_{\text{tot}}} \gamma \left(\boldsymbol{N}^{(j)}(t') - \boldsymbol{N}^{(i)}(t')\right)_x \\
&= \sum_j \lambda_{ij} \gamma \left(\boldsymbol{N}^{(j)}(t') - \boldsymbol{N}^{(i)}(t')\right)_x dt\,.
\end{aligned} \tag{3.30}$$

Applying the law of total variance, the variance of each entry of the sum is

$$\begin{aligned}
\operatorname{Var}\left[\left(\sum_{l=1}^{L}\widetilde{J}_l^{(i)}\right)_x\right] &= \mathbb{E}\left[\operatorname{Var}\left[\left(\sum_{l=1}^{L}\widetilde{J}_l^{(i)}\right)_x \,\middle|\, L\right]\right] + \operatorname{Var}\left[\mathbb{E}\left[\left(\sum_{l=1}^{L}\widetilde{J}_l^{(i)}\right)_x \,\middle|\, L\right]\right] \qquad (3.31)\\
&= \mathbb{E}\left[L\,\operatorname{Var}\left[\left(\widetilde{J}_l^{(i)}\right)_x\right]\right] + \operatorname{Var}\left[L\,\mathbb{E}\left[\left(\widetilde{J}_l^{(i)}\right)_x\right]\right]\\
&= \lambda_{\text{tot}}dt\,\operatorname{Var}\left[\left(\widetilde{J}_l^{(i)}\right)_x\right] + \lambda_{\text{tot}}dt\left(\mathbb{E}\left[\left(\widetilde{J}_l^{(i)}\right)_x\right]\right)^2\\
&= \lambda_{\text{tot}}dt\,\mathbb{E}\left[\left(\widetilde{J}_l^{(i)}\right)_x^2\right]\\
&= \lambda_{\text{tot}}dt\sum_j \frac{\lambda_{ij}}{\lambda_{\text{tot}}}\gamma^2\left(\boldsymbol{N}^{(j)}(t') - \boldsymbol{N}^{(i)}(t')\right)_x^2\\
&= \sum_j \lambda_{ij}\gamma^2\left(\boldsymbol{N}^{(j)}(t') - \boldsymbol{N}^{(i)}(t')\right)_x^2 dt\,.
\end{aligned}$$

Because $\lambda_{\text{tot}}\gamma$ is fixed, the interaction strength $\gamma \to 0$ as $\lambda_{\text{tot}} \to \infty$. As $\gamma \to 0$, the expectation (3.30) of each entry of (3.28) remains fixed, but the variance (3.31) of each entry decreases to 0. For sufficiently small γ, the effect of interactions on the microbiome abundance vector $\boldsymbol{N}^{(i)}(t)$ over the interval $[t', t'+dt]$ is

$$\left(\boldsymbol{N}^{(i)}(t'+dt) - \boldsymbol{N}^{(i)}(t')\right)_{\text{exch}} = \sum_j \lambda_{ij}\gamma\left(\boldsymbol{N}^{(j)}(t') - \boldsymbol{N}^{(i)}(t')\right)dt + O(dt^2) \qquad (3.32)$$

with arbitrarily high probability.

For sufficiently small γ, the change in the microbiome abundance vector $\boldsymbol{N}^{(i)}(t)$ over the interval $[t', t'+dt]$ is approximately equal to the sum of the effect (3.24) of local dynamics and the effect (3.32) of interactions. In Section 3.7.2, we show that

$$\begin{aligned}
\boldsymbol{N}^{(i)}(t'+dt) - \boldsymbol{N}^{(i)}(t') &= \left(\boldsymbol{N}^{(i)}(t'+dt) - \boldsymbol{N}^{(i)}(t')\right)_{\text{loc}} \qquad (3.33)\\
&\quad + \left(\boldsymbol{N}^{(i)}(t'+dt) - \boldsymbol{N}^{(i)}(t')\right)_{\text{exch}} + O(dt^2)\\
&= \left[g^{(i)}\left(\boldsymbol{N}^{(i)}(t')\right) + \sum_j \lambda_{ij}\gamma\left(\boldsymbol{N}^{(j)}(t') - \boldsymbol{N}^{(i)}(t')\right)\right]dt + O(dt^2)
\end{aligned}$$

with arbitrarily high probability. Therefore, $\boldsymbol{N}^{(i)}(t)$ is well-approximated by $\widetilde{\boldsymbol{N}}^{(i)}(t)$.

As an example of the HFLSA, consider a two-host system in which each host has local dynamics (3.12). Let $\boldsymbol{N}^{(1)}(0) = (2,2)$ and $\boldsymbol{N}^{(2)}(0) = (12,12)$. In Figure 3.8, we show

how the approximation improves as we increase the total-interaction-frequency parameter $\lambda_{\text{tot}} = \lambda_{12}$ and decrease the interaction strength γ for fixed $\lambda_{\text{tot}}\gamma = 8$. The error that we obtain by using the approximate microbiome abundance vector $\widetilde{\boldsymbol{N}}^{(i)}(t)$ appears to be largest near times that the ith abundance vector $\widetilde{\boldsymbol{N}}^{(i)}(t)$ transitions between different basins of attraction. However, for sufficiently large λ_{tot}, this error becomes arbitrarily small for all times $t \in [0, T]$ with arbitrarily high probability.

3.4.2 High-Frequency, Constant-Strength Approximation (HFCSA)

The HFCSA is accurate when interactions are very frequent and have constant strengths. In this regime, all microbiome abundance vectors converge rapidly to the mean microbiome abundance vector

$$\overline{\boldsymbol{N}}(t) = \frac{1}{|H|} \sum_{j=1}^{|H|} \boldsymbol{N}^{(j)}(t)\,. \tag{3.34}$$

Subsequently, these "synchronized" microbiome abundance vectors each follow the mean of their local dynamics (see (3.38) below).

Theorem 3 (High-Frequency, Constant-Strength Approximation Theorem)**.** *Fix the relative interaction-frequency parameters l_{ij}, the interaction strength $\gamma > 0$, and a time T. Suppose that each local-dynamics function $g^{(i)}$ is Lipschitz continuous and bounded (see Section 3.2.3). Let $\varepsilon \in (0, 1]$, $\delta > 0$, and $\eta > 0$ be arbitrary but fixed constants. For sufficiently large λ_{tot}, each host microbiome abundance vector $\boldsymbol{N}^{(i)}(t)$ satisfies*

$$\left\| \boldsymbol{N}^{(i)} - \widetilde{\boldsymbol{N}} \right\|_{L^\infty[\eta, T]} < \delta \tag{3.35}$$

with probability larger than $1 - \varepsilon$, where

$$\begin{aligned} \frac{d\widetilde{\boldsymbol{N}}}{dt} &= \frac{1}{|H|} \sum_{j=1}^{|H|} g^{(j)}\left(\widetilde{\boldsymbol{N}}\right)\,, \\ \widetilde{\boldsymbol{N}}(0) &= \overline{\boldsymbol{N}}(0)\,. \end{aligned} \tag{3.36}$$

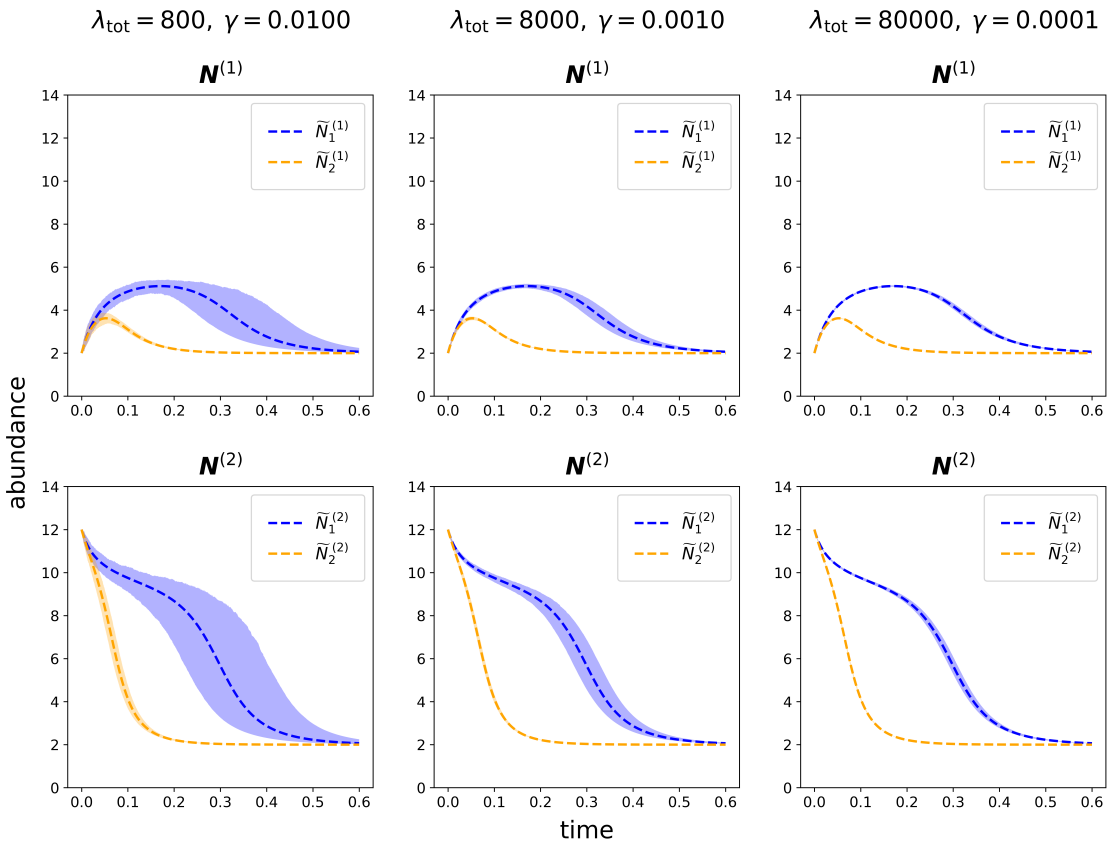


Figure 3.8: Numerical experiments for a two-host system in which each host has local dynamics (3.12). The three columns show experiments for different values of the total-interaction-frequency parameter λ_{tot} and the interaction strength γ for fixed $\lambda_{\text{tot}}\gamma = 8$. We show the microbiome abundances for $H^{(1)}$ in the first row and the microbiome abundances for $H^{(2)}$ in the second row. We run 500 simulations for each set of parameters. The highlighted region shows the range between the 5^{th} and 95^{th} percentiles of the simulated host abundances. The dashed curves show the HFLSAs for these experiments.

We provide key steps of the proof of Theorem 3 in this section. We give a full proof in Section 3.7.3.

When two hosts $H^{(i)}$ and $H^{(j)}$ interact at time t_I, they exchange portions of their microbiome as described in (3.4). This exchange causes no change in the sum

$$\begin{aligned}\boldsymbol{N}^{(i)}\left(t_I^+\right)+\boldsymbol{N}^{(j)}\left(t_I^+\right) &= (1-\gamma)\boldsymbol{N}^{(i)}\left(t_I^-\right)+\gamma\boldsymbol{N}^{(j)}\left(t_I^-\right)+(1-\gamma)\boldsymbol{N}^{(j)}\left(t_I^-\right)+\gamma\boldsymbol{N}^{(i)}\left(t_I^-\right) \\ &= \boldsymbol{N}^{(i)}\left(t_I^-\right)+\boldsymbol{N}^{(j)}\left(t_I^-\right)\ .\end{aligned} \tag{3.37}$$

Therefore, interactions do not cause any direct change in the mean microbiome abundance vector $\overline{\boldsymbol{N}}(t)$. The local dynamics drive the evolution

$$\frac{\mathrm{d}\overline{\boldsymbol{N}}}{\mathrm{d}t}=\frac{1}{|H|}\sum_{j=1}^{|H|}g^{(j)}\left(\boldsymbol{N}^{(j)}\right)\ . \tag{3.38}$$

For sufficiently large λ_{tot}, interactions occur on a much faster time scale than the local dynamics. Because of this separation of time scales, all microbiome abundance vectors $\boldsymbol{N}^{(i)}(t)$ converge rapidly, which entails that

$$\left\|\boldsymbol{N}^{(i)}-\overline{\boldsymbol{N}}\right\|_{L^\infty[\eta,T]}<\xi \tag{3.39}$$

with high probability. For fixed probability, a larger λ_{tot} allows a smaller bound ξ. For a sufficiently small bound ξ, each abundance vector $\boldsymbol{N}^{(i)}(t)$ is close enough to $\overline{\boldsymbol{N}}(t)$ so that $\overline{\boldsymbol{N}}(t)$ is well-approximated by the approximate microbiome abundance vector $\widetilde{\boldsymbol{N}}(t)$ (see (3.36)). For times $t\in[\eta,T]$, each microbiome abundance vector $\boldsymbol{N}^{(i)}(t)$ is very close to $\overline{\boldsymbol{N}}(t)$ and $\overline{\boldsymbol{N}}(t)$ is very close to $\widetilde{\boldsymbol{N}}(t)$. Therefore, on the interval $[\eta,T]$, each abundance vector $\boldsymbol{N}^{(i)}(t)$ is well-approximated by $\widetilde{\boldsymbol{N}}(t)$.

As an example of the HFCSA, consider a two-host system in which each host has local dynamics (3.12). Let $\boldsymbol{N}^{(1)}(0)=(2,2)$ and $\boldsymbol{N}^{(2)}(0)=(12,12)$. In Figure 3.9, we show how the approximation improves as we increase the total-interaction-frequency parameter $\lambda_{\mathrm{tot}}=\lambda_{12}$ for fixed interaction strength $\gamma=0.02$.

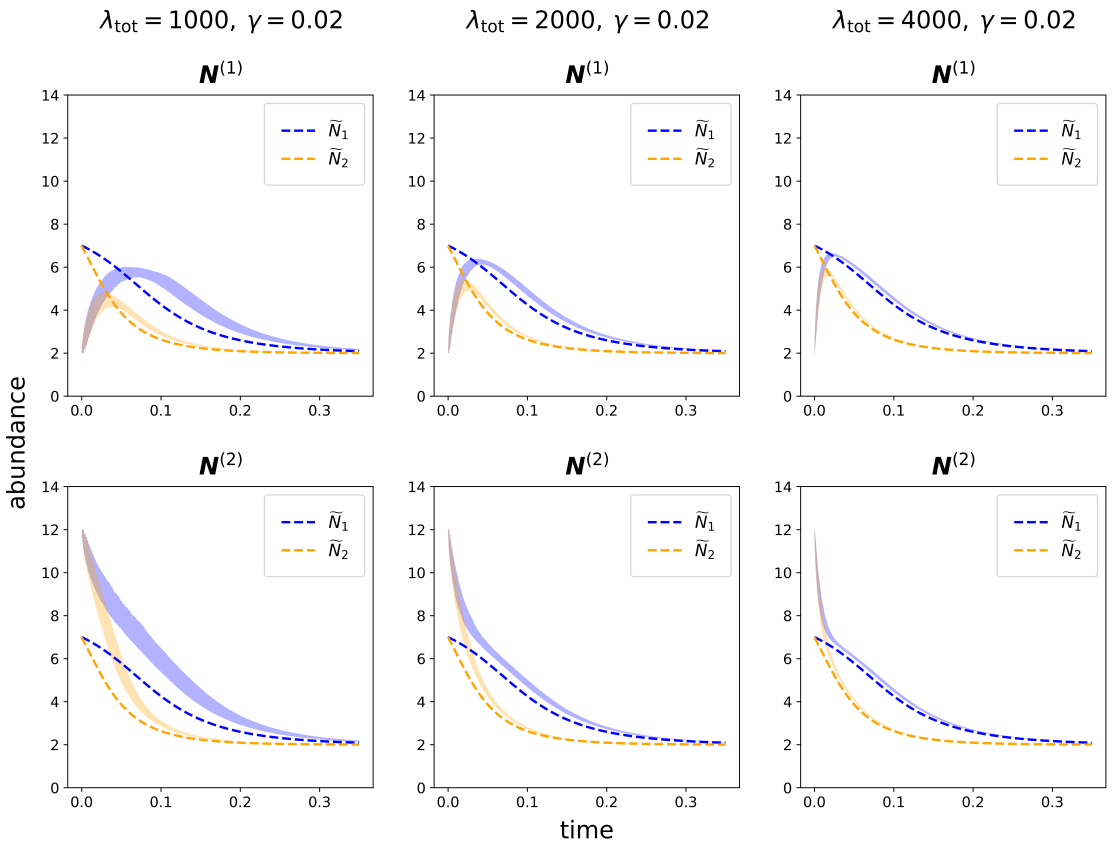


Figure 3.9: Numerical experiments for a two-host system in which each host has local dynamics (3.12). The three columns show experiments for different values of the total-interaction-frequency parameter λ_{tot} for fixed interaction strength $\gamma = 0.02$. We show the microbiome abundances for $H^{(1)}$ in the first row and the microbiome abundances for $H^{(2)}$ in the second row. We run 500 simulations for each set of parameters. The highlighted region shows the range between the 5^{th} and 95^{th} percentiles of the simulated host abundances. The dashed curves show the HFCSA for these experiments.

3.5 Numerical Experiments

In this section, we present simulations for a system of 10 hosts with local dynamics (3.12). We showed the interaction network for this system in Figure 3.1. This network has 25 edges, and we suppose that all relative interaction-frequency parameters λ_{ij} are equal. For each $\left(H^{(i)}, H^{(j)}\right) \in E$, the corresponding relative interaction-frequency parameter is $l_{ij} = 1/25$. We explore the accuracy of our three approximations for a range of values for the total-interaction-frequency parameter λ_{tot} and the interaction strength γ.

3.5.1 Pair Approximation for the LFA

For the LFA, the approximate basin probability tensor $\widetilde{\Psi}(t)$ (see (3.21)) has dimension $m_1 \times m_2 \times \cdots \times m_{|H|}$. In many circumstances, this tensor is too large to analyze directly. Therefore, we use a pair approximation [PG16, New18] of $\widetilde{\Psi}(t)$ for our calculations of the LFA in this subsection.

We approximate $\widetilde{\Psi}(t)$ by tracking the individual probabilities

$$\widetilde{\psi}_a^{(i)}(t) = \text{ probability that host } \boldsymbol{N}^{(i)}(t) \text{ is in basin } a \tag{3.40}$$

and the dyadic (i.e., pair) probabilities

$$\widetilde{\psi}_{ab}^{(ij)}(t) = \text{ probability that host } \boldsymbol{N}^{(i)}(t) \text{ is in basin } a \text{ and host } \boldsymbol{N}^{(j)}(t) \text{ is in basin } b\,. \tag{3.41}$$

To obtain a pair approximation, we also need to consider the triadic (i.e., triplet) probabilities

$$\begin{aligned}\widetilde{\psi}_{abc}^{(ijk)}(t) = \text{ probability that host } \boldsymbol{N}^{(i)}(t) \text{ is in basin } a, \text{ host } \boldsymbol{N}^{(j)}(t) \text{ is in basin } b, \\ \text{and host } \boldsymbol{N}^{(k)}(t) \text{ is in basin } c\,.\end{aligned} \tag{3.42}$$

In the LFA, if an interaction between hosts $H^{(i)}$ and $H^{(j)}$ causes their microbiome abundance vectors to move from basin d to basin a and from basin e to basin b, respectively, then

$\Phi^{(ij)}_{abde} = 1$. Otherwise, $\Phi^{(ij)}_{abde} = 0$. The probability that a given interaction is one between $H^{(i)}$ and $H^{(j)}$ is l_{ij}. Therefore, the change in $\widetilde{\boldsymbol{\psi}}^{(i)}_a(t)$ due to an interaction at time t_I is

$$\widetilde{\psi}^{(i)}_a(t_I^+) - \widetilde{\psi}^{(i)}_a(t_I^-) = \sum_{j,b,e} \sum_{d \neq a} l_{ij} \Phi^{(ij)}_{abde} \widetilde{\psi}^{(ij)}_{de}(t_I^-) - \sum_{j,b,e} \sum_{d \neq a} l_{ij} \Phi^{(ij)}_{deab} \widetilde{\psi}^{(ij)}_{ab}(t_I^-) \,. \tag{3.43}$$

Using frequency-scaled time, the expected number of system interactions during an interval $[t^*, t^* + dt^*]$ is dt^*. Therefore,

$$\frac{\mathrm{d}}{\mathrm{d}t^*} \widetilde{\psi}^{(i)}_a = \sum_{j,b,e} \sum_{d \neq a} l_{ij} \left[\Phi^{(ij)}_{abde} \widetilde{\psi}^{(ij)}_{de} - \Phi^{(ij)}_{dbae} \widetilde{\psi}^{(ij)}_{ae} \right] . \tag{3.44}$$

For the dyadic probabilities, we have

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t^*} \widetilde{\psi}^{(ij)}_{ab} = & \sum_{e \neq b, d \neq a} l_{ij} \left[\Phi^{(ij)}_{abde} \widetilde{\psi}^{(ij)}_{de} - \Phi^{(ij)}_{deab} \widetilde{\psi}^{(ij)}_{ab} \right] \\ & + \sum_{k,c,f} \sum_{d \neq a} l_{ik} \left[\Phi^{(ik)}_{acdf} \widetilde{\psi}^{(ijk)}_{dbf} - \Phi^{(ik)}_{dcaf} \widetilde{\psi}^{(ijk)}_{abf} \right] \\ & + \sum_{k,c,f} \sum_{e \neq b} l_{jk} \left[\Phi^{(jk)}_{bcef} \widetilde{\psi}^{(ijk)}_{aef} - \Phi^{(jk)}_{ecbf} \widetilde{\psi}^{(ijk)}_{abf} \right] . \end{aligned} \tag{3.45}$$

The right-hand sides in (3.44) and (3.45) are exact expressions. We form an approximation by replacing the triadic probabilities in (3.45) with combinations of dyadic probabilities. In the derivatives of $\widetilde{\boldsymbol{\psi}}^{(ij)}(t^*)$, when considering the impact of interactions between hosts $H^{(i)}$ and $H^{(k)}$, we use the approximation

$$\widetilde{\psi}^{(ijk)}_{abc}(t^*) \approx \frac{\widetilde{\psi}^{(ij)}_{ab}(t^*) \widetilde{\psi}^{(ik)}_{ac}(t^*)}{\widetilde{\psi}^{(i)}_a(t^*)} \,, \tag{3.46}$$

which assumes that $\widetilde{\psi}^{(j)}_b$ and $\widetilde{\psi}^{(k)}_c$ are independent. Inserting the approximation (3.46) into (3.45) gives

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t^*} \widetilde{\psi}^{(ij)}_{ab} \approx & \sum_{e \neq b, d \neq a} l_{ij} \left[\Phi^{(ij)}_{abde} \widetilde{\psi}^{(ij)}_{de} - \Phi^{(ij)}_{deab} \widetilde{\psi}^{(ij)}_{ab} \right] \\ & + \sum_{k,c,f} \sum_{d \neq a} l_{ik} \left[\Phi^{(ik)}_{acdf} \frac{\widetilde{\psi}^{(ij)}_{db} \widetilde{\psi}^{(ik)}_{df}}{\widetilde{\psi}^{(i)}_d} - \Phi^{(ik)}_{dcaf} \frac{\widetilde{\psi}^{(ij)}_{ab} \widetilde{\psi}^{(ik)}_{af}}{\widetilde{\psi}^{(i)}_a} \right] \\ & + \sum_{k,c,f} \sum_{e \neq b} l_{jk} \left[\Phi^{(jk)}_{bcef} \frac{\widetilde{\psi}^{(ij)}_{ae} \widetilde{\psi}^{(jk)}_{ef}}{\widetilde{\psi}^{(j)}_e} - \Phi^{(jk)}_{ecbf} \frac{\widetilde{\psi}^{(ij)}_{ab} \widetilde{\psi}^{(jk)}_{bf}}{\widetilde{\psi}^{(j)}_b} \right] . \end{aligned} \tag{3.47}$$

Equations (3.44) and (3.47) constitute a pair approximation of the evolution of $\widetilde{\Psi}(t^*)$. In our simulations (see Section 3.5.2), we compare the individual probabilities $\widetilde{\psi}_a^{(i)}(t^*)$ from our pair approximation (3.44, 3.47) with the fraction $\psi_a^{(i),\,\mathrm{sim}}(t^*)$ of simulations in which $\boldsymbol{N}^{(i)}(t^*)$ is in basin of attraction a.

3.5.2 Simulations for the Low-Frequency Approximation

In this subsection, we show numerical results for the LFA (3.21). We use the pair approximation (3.44, 3.47) to determine the individual probabilities $\widetilde{\psi}_a^{(i)}(t^*)$). As we described in Section 3.5.1, an individual probability $\widetilde{\psi}_a^{(i)}(t^*)$ is an approximation of the probability from the LFA that $\boldsymbol{N}^{(i)}(t^*)$ is in basin of attraction a. We compare these individual probabilities to the fraction $\psi_a^{(i),\,\mathrm{sim}}(t^*)$ of simulations of (3.4, 3.12) in which $\boldsymbol{N}^{(i)}(t^*)$ is in basin of attraction a. We perform these approximations and simulations for 59 linearly spaced interaction strengths $\gamma \in [0, 0.5]$ and 13 logarithmically spaced values of the total-interaction-frequency parameter $\lambda_{\mathrm{tot}} \in [2.5 \times 10^{-2}, 2.5 \times 10^{4}]$. Because $\mathcal{B} = \{0.1, 0.4\}$ for this system, the LFA is not valid when $\gamma = 0.1$ or $\gamma = 0.4$, so we exclude these values of γ from our simulations. Therefore, we perform simulations for values of $\gamma \in [0, 0.5]$ that are multiples of $\frac{0.5}{60}$ (except $\gamma = 0.1$ and $\gamma = 0.4$). For each pair of γ and λ_{tot} values, we select a random four-dimensional (4D) vector $\widetilde{\boldsymbol{\psi}}^{(i)}(0)$ from the Dirichlet distribution $\mathrm{Dir}(1, 1, 1, 1)$ [Mac05] for each host. The entries of each of these 4D vectors sum to 1. For each simulation, we set each $\boldsymbol{N}^{(i)}(0)$ to be the stable equilibrium point in one of the basins of attraction. For each basin of attraction a, the initial microbiome abundance vector $\boldsymbol{N}^{(i)}(0)$ is in that basin of attraction with probability $\widetilde{\psi}_a^{(i)}(0)$.

We perform 1000 simulations of (3.4, 3.12) for each pair of γ and λ_{tot} values, and we calculate the fraction $\psi_a^{(i),\,\mathrm{sim}}(t^*)$ of the simulations in which $\boldsymbol{N}^{(i)}(t^*)$ is in basin of attraction a. We compare $\psi_a^{(i),\,\mathrm{sim}}(t^*)$ to $\widetilde{\boldsymbol{\psi}}^{(i)}(t^*)$, which we calculate using the pair approximation (3.44, 3.47). We calculate $\boldsymbol{\psi}^{(i),\,\mathrm{sim}}(t^*)$ and $\widetilde{\boldsymbol{\psi}}^{(i)}(t^*)$ for 1001 evenly spaced frequency-scaled times

$t_k^* = k/500$ in the interval $[0, 2]$. In Figure 3.10, we plot the error

$$\text{Error} = \frac{2}{1001} \sum_{k=1}^{1001} \sqrt{\sum_{i=1}^{10} \sum_{a=1}^{4} \left(\psi_a^{(i),\,\text{sim}}(t_k^*) - \widetilde{\psi}_a^{(i)}(t_k^*) \right)^2} \tag{3.48}$$

for each pair of γ and λ_{tot} values. The error (3.48) is a discrete approximation of the norm $\left\| \boldsymbol{\psi}^{(i),\,\text{sim}} - \widetilde{\boldsymbol{\psi}}^{(i)} \right\|_{L^2[0,T^*]}$.

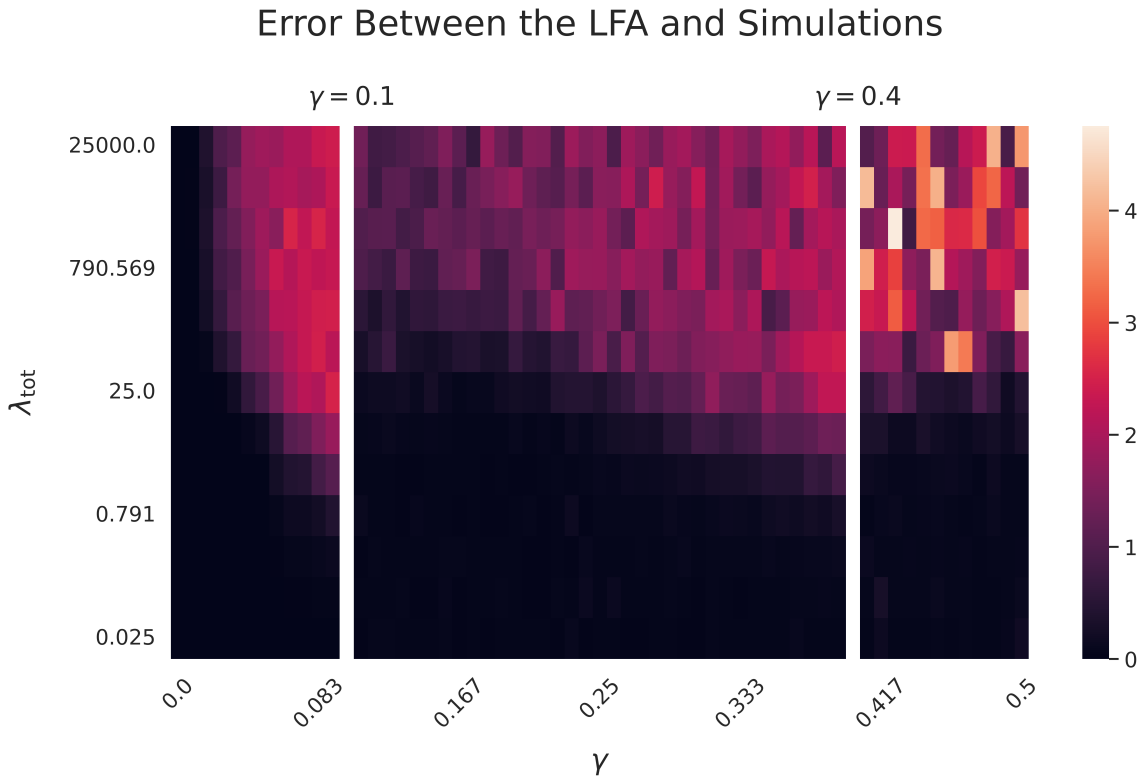


Figure 3.10: The error (3.48) between the LFA (see (3.21)) versus means of 1000 simulations of (3.4, 3.12) for each pair of the interaction strength γ and the total-interaction-frequency parameter λ_{tot}. We plot γ on a linear scale and λ_{tot} on a logarithmic scale. We do not plot errors for $\gamma = 0.1$ and $\gamma = 0.4$ because the LFA is not valid for these values.

The LFA is most accurate when the total-interaction-frequency parameter λ_{tot} is small. It is also better when the interaction strength γ is not near 0.1 or 0.4. There are two types of errors in the LFA. The first type of error arises when repeated interactions occur

in sufficiently quick succession to yield a transition that the LFA misses. For example, for $\gamma \approx 0.342$, the LFA does not predict that the $\boldsymbol{N}^{(i)}(t)$ can move from basin 2 to basin 1. For sufficiently large values of λ_{tot}, repeated interactions in short succession are common (and not merely possible), which causes the LFA to overestimate the probability that a host is in basin 2 and underestimate the probability that a host is in basin 1. In Figure 3.11, we illustrate this type of error for $\gamma \approx 0.342$ and several values of λ_{tot}.

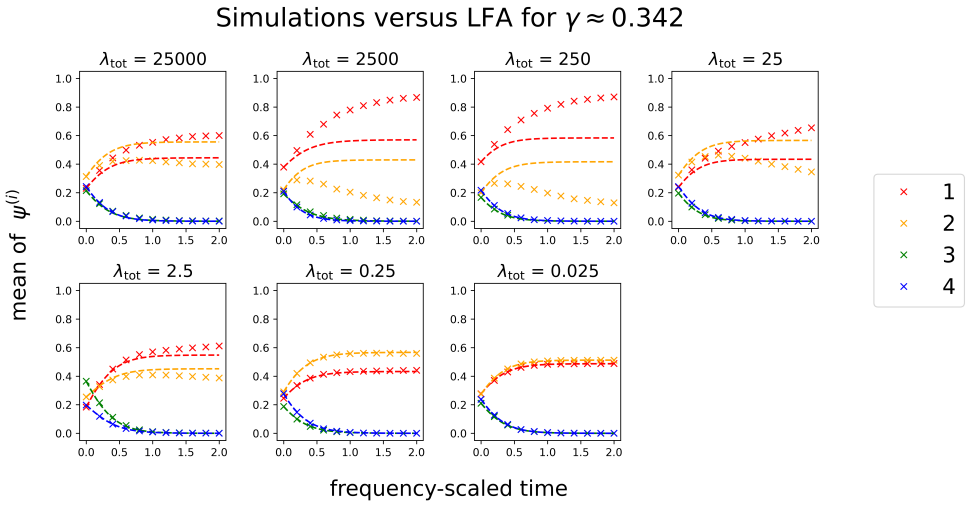


Figure 3.11: The means of the simulated probabilities $\psi_a^{(i),\,\text{sim}}(t^*)$ over all hosts for interaction strength $\gamma \approx 0.342$ and several values of the total-interaction-frequency parameter λ_{tot}. The dashed curves indicate the LFA approximation of the mean of the probabilities $\widetilde{\boldsymbol{\psi}}^{(i)}(t^*)$ over all hosts.

The second type of error arises when repeated interactions in sufficiently quick succession cause the LFA to overestimate the impact of the second and subsequent interactions. As an example, for the interaction strength $\gamma \approx 0.433$, the LFA predicts that $\boldsymbol{N}^{(j)}(t)$ will be in basin 1 after an interaction whenever $H^{(j)}$ interacts with $H^{(i)}$ and $\boldsymbol{N}^{(i)}(t)$ is in basin 1 before the interaction. This prediction arises because the LFA assumes that $\boldsymbol{N}^{(i)} = (2, 2)$ before this interaction. However, if $H^{(i)}$ recently interacted with a different host, then $\boldsymbol{N}^{(i)}(t)$ may be in basin 1 but not sufficiently close to the equilibrium point $(2, 2)$ to drive a transition

to the basin of attraction of $\boldsymbol{N}^{(j)}(t)$. Consequently, the LFA overestimates the probability that a host is in basin 1. In Figure 3.12, we illustrate this type of error for $\gamma \approx 0.433$ and several values of λ_{tot}.

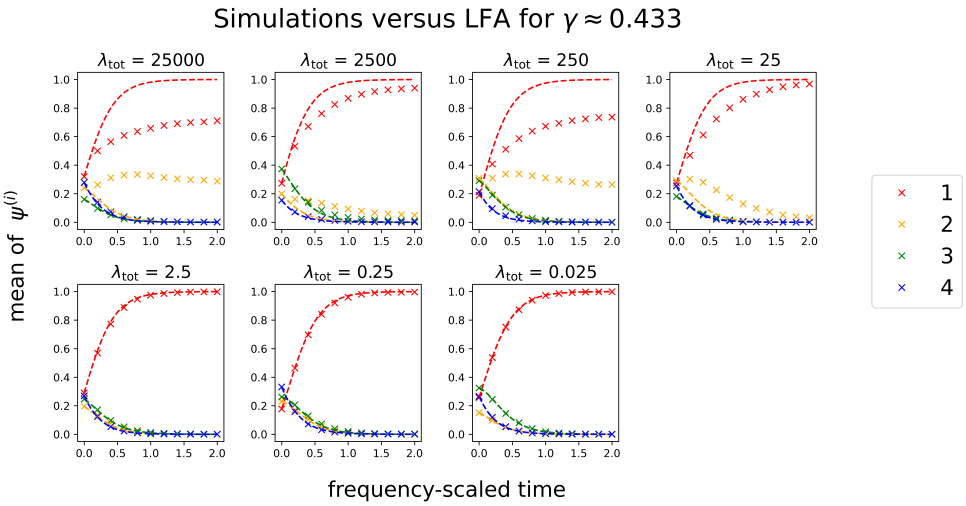


Figure 3.12: The mean of the simulated probabilities $\psi_a^{(i),\,\text{sim}}(t^*)$ over all hosts for interaction strength $\gamma \approx 0.433$ and several values of the total-interaction-frequency parameter λ_{tot}. The dashed curves indicate the LFA approximation of the mean of the probabilities $\widetilde{\boldsymbol{\psi}}^{(i)}(t^*)$ over all hosts.

3.5.3 Simulations for the High-Frequency Approximations

In this subsection, we show numerical results for the HFLSA (3.23) and the HFCSA (3.36). We compare the approximate microbiome abundance vectors from these two approximations to simulations of (3.4, 3.12).

To evaluate the accuracy of the approximate microbiome abundance vectors $\widetilde{\boldsymbol{N}}^{(i)}(t)$ that we obtain from the HFLSA (3.23), we compare them to the microbiome abundance vectors $\boldsymbol{N}^{(i)}(t)$ from (3.4, 3.12). We perform these simulations for 13 logarithmically spaced values of the total-interaction-frequency parameter $\lambda_{\text{tot}} \in [25, 2500]$ and 75 linearly spaced values of $\lambda_{\text{tot}}\gamma \in [0.04, 0.3]$. We use the product $\lambda_{\text{tot}}\gamma$ as a parameter instead of the interaction

strength γ on its own to illustrate the improvement of the HFLSA as we increase λ_{tot} for fixed $\lambda_{\text{tot}}\gamma$. For each pair of λ_{tot} and $\lambda_{\text{tot}}\gamma$ values, we perform 1000 simulations over the time interval $[0,1]$ with initial conditions

$$\begin{aligned}
&\boldsymbol{N}^{(1)}(0) = (12,12)\,,\ \boldsymbol{N}^{(2)}(0) = (2,2)\,,\ \boldsymbol{N}^{(3)}(0) = (12,2)\,, \\
&\boldsymbol{N}^{(4)}(0) = (2,2)\,,\ \boldsymbol{N}^{(5)}(0) = (12,12)\,,\ \boldsymbol{N}^{(6)}(0) = (12,12)\,, \\
&\boldsymbol{N}^{(7)}(0) = (2,2)\,,\ \boldsymbol{N}^{(8)}(0) = (12,2)\,,\ \boldsymbol{N}^{(9)}(0) = (2,12)\,,\ \boldsymbol{N}^{(10)}(0) = (2,12)\,.
\end{aligned} \tag{3.49}$$

The microbiome abundance vector $\boldsymbol{N}^{(i),l}(t)$ is the lth simulated microbiome abundance vector for host $H^{(i)}$. For each of these simulations, we calculate the simulated microbiome abundance vectors $\boldsymbol{N}^{(i),l}(t)$ for 101 evenly spaced times $t_k = k/100$. We compare these simulations to the approximate microbiome abundance vector $\widetilde{\boldsymbol{N}}^{(i)}(t)$ from the HFLSA (3.23). In Figure 3.13, we plot the error

$$\text{Error} = \frac{1}{101 \times 1000} \sum_{l=1}^{1000} \sum_{k=1}^{101} \sqrt{\sum_{i=1}^{10} \sum_{a=1}^{2} \left(\boldsymbol{N}_a^{(i),l}(t_k) - \widetilde{\boldsymbol{N}}_a^{(i)}(t_k) \right)^2} \tag{3.50}$$

for each pair of λ_{tot} and $\lambda_{\text{tot}}\gamma$ values. The error (3.50) is a discrete approximation of the mean of $\left\| \boldsymbol{N}_a^{(i),l}(t_k) - \widetilde{\boldsymbol{N}}_a^{(i)}(t_k) \right\|_{L^2[0,1]}$ over all simulations. For any fixed value of $\lambda_{\text{tot}}\gamma$, the HFLSA is more accurate for larger λ_{tot}. However, for a fixed value of λ_{tot}, the error depends significantly on the value of $\lambda_{\text{tot}}\gamma$. Each value of $\lambda_{\text{tot}}\gamma$ yields a set $\left\{ \widetilde{\boldsymbol{N}}^{(i)}(1) \right\}$ of final approximate microbiome abundance vectors. For all but a finite set of values of $\lambda_{\text{tot}}\gamma$, each final approximate microbiome abundance vector $\widetilde{\boldsymbol{N}}^{(i)}(1)$ changes continuously with $\lambda_{\text{tot}}\gamma$. However, there are a finite number of $\lambda_{\text{tot}}\gamma$ values for which some final approximate microbiome abundance vector $\widetilde{\boldsymbol{N}}^{(i)}(1)$ has a discontinuous jump. For the initial set of microbiome abundance vectors (3.49), these discontinuous jumps occur at

$$\lambda_{\text{tot}}\gamma \in \{0.0866, 0.1069, 0.1160, 0.1161, 0.1416, 1.6432, 1.7174, 1.7425, 1.8187, 1.8363\}\,. \tag{3.51}$$

The regions in which the HFLSA performs worst in our simulations are near $\lambda_{\text{tot}}\gamma \approx 0.1$ and $\lambda_{\text{tot}}\gamma \approx 1.8$, which are very close to several of the values in (3.51).

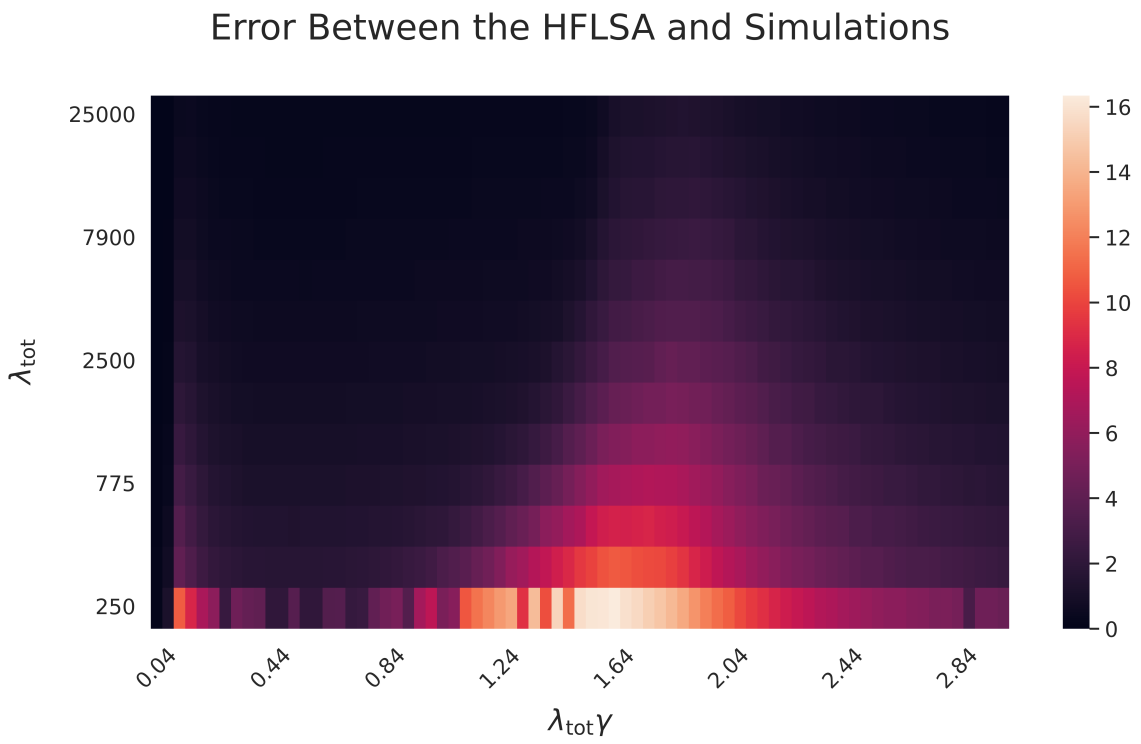


Figure 3.13: The mean error (3.50) between the approximate microbiome abundance vectors $\left\{\widetilde{\boldsymbol{N}}^{(i)}(t)\right\}$ from the HFLSA (see (3.23)) and the microbiome abundance vectors $\left\{\boldsymbol{N}^{(i)}(t)\right\}$ for 1000 simulations of (3.4, 3.12). We plot λ_{tot} on a logarithmic scale and plot $\lambda_{\text{tot}}\gamma$ on a linear scale.

For the HFCSA (3.36), we perform simulations for 61 linearly spaced values of the interaction strength $\gamma \in [0, 0.5]$ and 13 logarithmically spaced values of the total-interaction-frequency parameter $\lambda_{\text{tot}} \in [2.5 \times 10^{-1}, 2.5 \times 10^{4}]$. We use the same initial conditions (3.49) as in our HFLSA simulations. We also again perform 1000 simulations and evaluate each simulated microbiome abundance vector $\boldsymbol{N}^{(i),l}(t)$ for 101 evenly spaced times $t_k = k/100$ on the time interval $[0, 1]$. In Figure 3.14, we show the error (3.50) for each pair of γ and λ_{tot} values. The approximation is accurate for sufficiently large λ_{tot}. In general, a larger γ increases the rate at which each microbiome abundance vector $\boldsymbol{N}^{(i)}(t)$ converges to the mean microbiome abundance vector $\overline{\boldsymbol{N}}(t)$. Consequently, the HFCSA yields a better approximation of $\boldsymbol{N}^{(i)}(t)$ for larger values of γ.

We are currently working on heuristic approximations and experiments for the convergence rates of the microbiome abundance vectors $\boldsymbol{N}^{(i)}(t)$.

3.6 Conclusions and Discussion

3.6.1 Summary

We developed a novel framework to model the microbiome dynamics of living hosts that incorporates both the local dynamics within an environment and exchanges of microbiomes between environments. Our framework extends existing metacommunity theory by accounting for the discrete nature of host interactions. Unlike classical mass-effects models, our framework incorporates two distinct parameters that control interaction frequencies and interaction strength. Using both analytical approximations and numerical computations, we demonstrated that both parameters are necessary to determine microbiome dynamics.

We developed approximations in three parameter regions, and we proved their accuracy in those regions. Our low-frequency approximation (LFA) gives a good approximation of the microbiome dynamics when local dynamics are much faster than host interactions. Our high-frequency, low-strength approximation (HFLSA) encodes the dynamics of a sys-

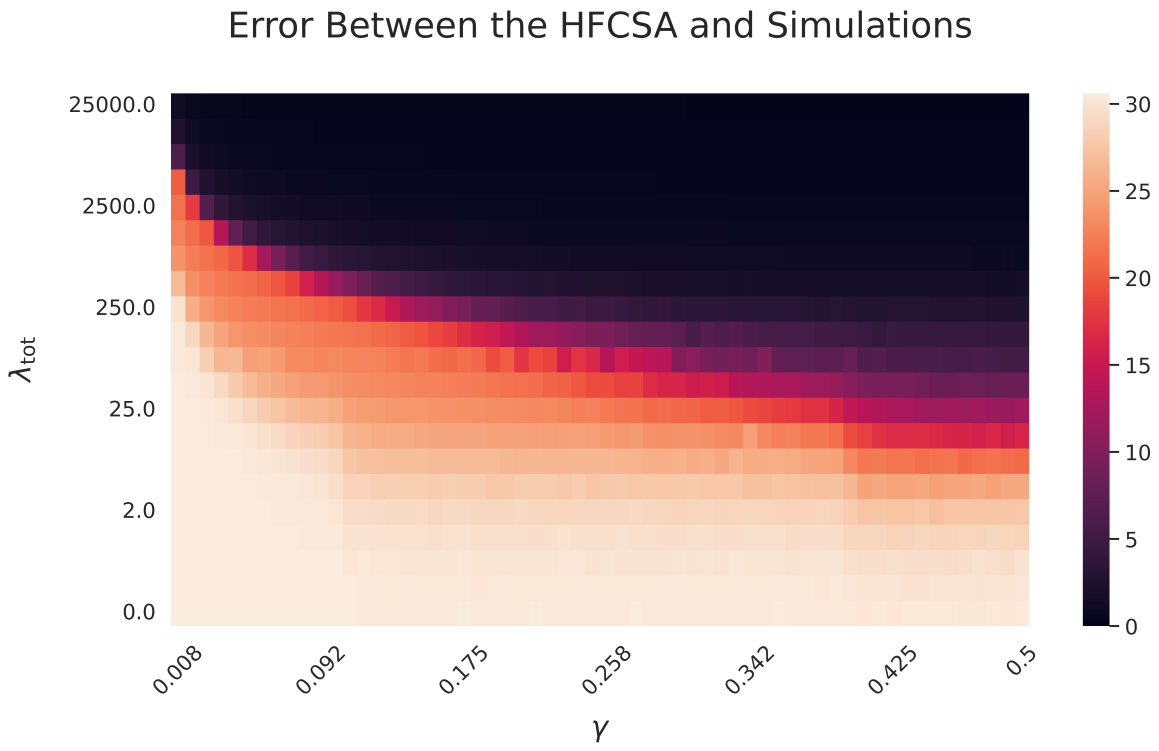


Figure 3.14: The mean error (3.50) between the approximate microbiome abundance vectors $\left\{\widetilde{\boldsymbol{N}}^{(i)}(t)\right\}$ from the HFCSA (see (3.36)) and the microbiome abundance vectors $\left\{\boldsymbol{N}^{(i)}(t)\right\}$ for 1000 simulations of (3.4, 3.12). We plot λ_{tot} on a logarithmic scale and plot γ on a linear scale.

tem when interactions are frequent but weak, resulting in a model with the same form as the mass-effects model (3.3). Finally, our high-frequency, constant-strength approximation (HFCSA) accurately predicts the rapid convergence of all hosts' microbiome dynamics when interactions are frequent and have constant interaction strength γ. We validated each of these approximations through numerical experiments on an illustrative model of microbiome dynamics for a range of parameter values.

A qualitative example of dynamics in our model involves the probability that all microbiome abundance vectors converge in some time interval. This probability depends both the interaction-frequency parameters and on the interaction strength. Using the LFA, we showed for sufficiently small interaction-frequency parameters that the convergence probability depends on whether the interaction strength γ is large enough that a single interaction between two hosts places their microbiome abundance vectors in the same basin of attraction. By contrast, using the HFLSA, we showed for sufficiently large interaction-frequency parameters that the convergence probability depends on the product $\lambda_{\text{tot}}\gamma$ of the total-interaction-frequency parameter λ_{tot} and the interaction strength γ. For intermediate values of the interaction-frequency parameters, we used numerical simulations of models in our framework to examine convergence probabilities.

3.6.2 Outlook

Our modeling framework provides a foundation for many promising future research directions in microbiome dynamics. In our framework's current form, one can use it to study the effects of host interactions in many ecological models of local dynamics. One can also use our framework to study the impact of the structure of interaction networks on microbiome dynamics.

There are many possible extensions of our modeling framework. For example, we considered a homogeneous interaction strength γ for simplicity. However, one can allow each pair of hosts to have heterogeneous interaction strengths γ_{ij}. Additionally, hosts can exchange

different microbe species with different exchange strengths, so one can also consider interaction strengths γ_{ijk} that encode the exchange of different species k when hosts $H^{(i)}$ and $H^{(j)}$ interact. This seems helpful when using consumer–resource models for the local dynamics. In this case, it also may be desirable to separately encode the strengths of microbiome exchange and resource exchange. Another way to extend our framework is to relax our assumption of instantaneous microbiome exchange by instead employing rapid but continuous functions.

Additionally, in our framework, our LFA assumes that the attractors of each local-dynamics function $g^{(i)}$ consist of a finite set of stable equilibrium points. We believe that it is possible to extend the LFA to systems in which the $g^{(i)}$ have more complicated attractors, such as limit cycles and chaotic attractors. We also believe that it is possible to generalize the HFLSA and HFCSA to systems in which the times between consecutive interactions for pairs of adjacent hosts follow a distribution other than an exponential distribution.

3.7 Proofs of our Approximations

3.7.1 Proof of Low-Frequency Approximation Theorem

In this appendix, we prove the LFA Theorem (see Theorem 1).

Theorem 1 (Low-Frequency-Approximation Theorem)**.** *Suppose that the attractors of each host's local dynamics consist of a finite set of stable equilibrium points at which the local-dynamics function $g^{(i)}$ is inward pointing, and let each $g^{(i)}$ be continuous and bounded (see Section 3.2.3). Fix $\gamma \notin \mathcal{B}$, all l_{ij}, and a frequency-scaled time T^*. As $\lambda_{\text{tot}} \to 0$, the basin probability tensor $\Psi(t^*)$ converges uniformly to $\widetilde{\Psi}(t^*)$ on $[0, T^*]$, where*

$$\frac{d}{dt^*}\widetilde{\Psi}_{b_1,\ldots,b_{|H|}}(t^*) = \Phi_{b_1,\ldots,b_{|H|},a_1,\ldots,a_{|H|}}\widetilde{\Psi}_{a_1,\ldots,a_{|H|}}(t^*) - \widetilde{\Psi}_{b_1,\ldots,b_{|H|}}(t^*)\,, \tag{3.21}$$

$$\widetilde{\Psi}(0) = \Psi(0)\,.$$

Proof. Each $g^{(i)}$ is bounded (see Section 3.2.3) so, there exists a constant M such that each

entry of $\boldsymbol{N}^{(i)}(t)$ is nonnegative and

$$\left\|\boldsymbol{N}^{(i)}(t)\right\|_{\infty} \leq M \tag{3.52}$$

for each microbiome abundance vector $\boldsymbol{N}^{(i)}(t)$ and all times $t \geq 0$.

Consider an arbitrary but fixed $\varepsilon > 0$. We will show that

$$\left\|\Psi(t^*) - \widetilde{\Psi}(t^*)\right\|_{\infty} < \varepsilon \tag{3.53}$$

for sufficiently small total-interaction-frequency parameter λ_{tot} and all frequency-scaled times $t^* \in [0, T^*]$.

For each i, let $\mathcal{A}_i$ be the set of stable equilibrium points of the local dynamics of host $H^{(i)}$. Suppose that adjacent hosts $H^{(i)}$ and $H^{(j)}$ interact at time t_I and that $\boldsymbol{N}^{(i)}\left(t_I^-\right)$ and $\boldsymbol{N}^{(j)}\left(t_I^-\right)$ are at stable equilibrium points $\boldsymbol{a}^{(i)} \in \mathcal{A}_i$ and $\boldsymbol{a}^{(j)} \in \mathcal{A}_j$, respectively. For $\boldsymbol{x}, \boldsymbol{y} \in [0, M]^n$, let

$$\mathcal{X}(\boldsymbol{x}, \boldsymbol{y}) = (1 - \gamma)\boldsymbol{x} + \gamma \boldsymbol{y}\,. \tag{3.54}$$

After the interaction, $\boldsymbol{N}^{(i)}\left(t_I^+\right) = \mathcal{X}\left(\boldsymbol{a}^{(i)}, \boldsymbol{a}^{(j)}\right)$ and $\boldsymbol{N}^{(j)}\left(t_I^+\right) = \mathcal{X}\left(\boldsymbol{a}^{(j)}, \boldsymbol{a}^{(i)}\right)$. Because $\gamma \notin \mathcal{B}$, it follows that $\mathcal{X}\left(\boldsymbol{a}^{(i)}, \boldsymbol{a}^{(j)}\right)$ and $\mathcal{X}\left(\boldsymbol{a}^{(j)}, \boldsymbol{a}^{(i)}\right)$ are in some basins of attraction b_i and b_j for the stable equilibrium points $\boldsymbol{b}^{(i)} \in \mathcal{A}_i$ and $\boldsymbol{b}^{(j)} \in \mathcal{A}_j$, respectively.

Let

$$B(\boldsymbol{x}, \delta) = \{\boldsymbol{y} \mid \|\boldsymbol{y} - \boldsymbol{x}\|_2 < \delta \text{ and } \boldsymbol{y} \in [0, M]^n\}\,, \tag{3.55}$$

$$\overline{B}(\boldsymbol{x}, \delta) = \{\boldsymbol{y} \mid \|\boldsymbol{y} - \boldsymbol{x}\|_2 \leq \delta \text{ and } \boldsymbol{y} \in [0, M]^n\}\,. \tag{3.56}$$

The basins of attraction of stable equilibrium points are open sets, so there exists $\delta(\boldsymbol{a}^{(i)}, \boldsymbol{a}^{(j)}) = \delta(\boldsymbol{a}^{(j)}, \boldsymbol{a}^{(i)})$ such that

$$\begin{aligned} B\left(\mathcal{X}\left(\boldsymbol{a}^{(i)}, \boldsymbol{a}^{(j)}\right), \delta(\boldsymbol{a}^{(i)}, \boldsymbol{a}^{(j)})\right) &\subseteq b_i\,, \\ B\left(\mathcal{X}\left(\boldsymbol{a}^{(j)}, \boldsymbol{a}^{(i)}\right), \delta(\boldsymbol{a}^{(i)}, \boldsymbol{a}^{(j)})\right) &\subseteq b_j\,. \end{aligned} \tag{3.57}$$

All $\mathcal{A}_i$ are finite, so there are minima

$$\delta_{ij} = \min_{\boldsymbol{a}^{(i)} \in \mathcal{A}_i} \min_{\boldsymbol{a}^{(j)} \in \mathcal{A}_j} \delta(\boldsymbol{a}^{(i)}, \boldsymbol{a}^{(j)}) \,. \tag{3.58}$$

Let the flow $\boldsymbol{X}^{(i)}(t, \boldsymbol{x})$ be the solution of

$$\begin{aligned} \frac{\partial \boldsymbol{X}^{(i)}}{\partial t}(t, \boldsymbol{x}) &= g^{(i)} \left(\boldsymbol{X}^{(i)}(t, \boldsymbol{x}) \right) \,, \\ \boldsymbol{X}^{(i)}(0, \boldsymbol{x}) &= \boldsymbol{x} \,. \end{aligned} \tag{3.59}$$

For each local-dynamics function $g^{(i)}$, each $\boldsymbol{a}^{(i)} \in \mathcal{A}_i$ is a stable equilibrium point at which $g^{(i)}$ is inward pointing. Therefore, there exists $\delta\left(\boldsymbol{a}^{(i)}\right)$ such that $g^{(i)}(\boldsymbol{y}) \cdot \left(\boldsymbol{a}^{(i)} - \boldsymbol{y}\right) > 0$ for $\boldsymbol{y} \in \overline{B}\left(\boldsymbol{a}^{(i)}, \delta\left(\boldsymbol{a}^{(i)}\right)\right)$. If $\boldsymbol{x}$ is in the basin of attraction of $\boldsymbol{a}^{(i)}$, then $\boldsymbol{X}(t_a, \boldsymbol{x}) \in B\left(\boldsymbol{a}^{(i)}, \delta\left(\boldsymbol{a}^{(i)}\right)\right)$ for some time t_a. We then have

$$\begin{aligned} \left[\frac{\partial}{\partial t} \left\| \boldsymbol{X}^{(i)} - \boldsymbol{a}^{(i)} \right\|_2^2 \right] (t_a, \boldsymbol{x}) &= \left[2 \frac{\partial \boldsymbol{X}^{(i)}}{\partial t} \cdot \left(\boldsymbol{X}^{(i)} - \boldsymbol{a}^{(i)} \right) \right] (t_a, \boldsymbol{x}) \,, \\ &= -2 g^{(i)} \left(\boldsymbol{X}^{(i)}(t_a, \boldsymbol{x}) \right) \cdot \left(\boldsymbol{a}^{(i)} - \left(\boldsymbol{X}^{(i)}(t_a, \boldsymbol{x}) \right) \right) < 0 \,. \end{aligned} \tag{3.60}$$

Consequently, $\left\| \boldsymbol{X}^{(i)}(t, \boldsymbol{x}) - \boldsymbol{a}^{(i)} \right\|_2$ is monotonically decreasing for $t \geq t_a$. There are only finitely many hosts, so there exists

$$\begin{aligned} \delta &= \frac{1}{2} \min \left\{ \Delta_1 \cup \Delta_2 \right\} \,, \\ \Delta_1 &= \bigcup_i \bigcup_{\boldsymbol{a}^{(i)} \in \mathcal{A}_i} \left\{ \delta\left(\boldsymbol{a}^{(i)}\right) \right\} \,, \\ \Delta_2 &= \bigcup_{\{i,j \;\mid\; \left(H^{(i)}, H^{(j)}\right) \in E\}} \{\delta_{ij}\} \,. \end{aligned} \tag{3.61}$$

For any $\boldsymbol{x}, \boldsymbol{y} \in [0, M]^n$, let

$$\mathcal{F}(\boldsymbol{x}, \boldsymbol{y}) = \overline{B}\left(\mathcal{X}(\boldsymbol{x}, \boldsymbol{y}), \delta \right) \,. \tag{3.62}$$

For each pair of adjacent hosts $H^{(i)}$ and $H^{(j)}$ and each $\boldsymbol{a}^{(i)} \in \mathcal{A}_i$ and $\boldsymbol{a}^{(j)} \in \mathcal{A}_j$, we have

$$\begin{aligned} \mathcal{F}\left(\boldsymbol{a}^{(i)}, \boldsymbol{a}^{(j)}\right) &= \overline{B}\left(\mathcal{X}\left(\boldsymbol{a}^{(i)}, \boldsymbol{a}^{(j)}\right), \delta \right) \subset B\left(\mathcal{X}\left(\boldsymbol{a}^{(i)}, \boldsymbol{a}^{(j)}\right), \delta(\boldsymbol{a}^{(i)}, \boldsymbol{a}^{(j)}) \right) \subseteq b_i \,, \\ \mathcal{F}\left(\boldsymbol{a}^{(j)}, \boldsymbol{a}^{(i)}\right) &= \overline{B}\left(\mathcal{X}\left(\boldsymbol{a}^{(j)}, \boldsymbol{a}^{(i)}\right), \delta \right) \subset B\left(\mathcal{X}\left(\boldsymbol{a}^{(j)}, \boldsymbol{a}^{(i)}\right), \delta(\boldsymbol{a}^{(i)}, \boldsymbol{a}^{(j)}) \right) \subseteq b_j \end{aligned} \tag{3.63}$$

for some basins of attraction b_i and b_j of the local dynamics of hosts $H^{(i)}$ and $H^{(j)}$, respectively.

Recall that $\mathcal{U}^{(i)} \in [0, M]^n$ is the set of points that are not in the basin of attraction of some stable equilibrium point of the local dynamics of host $H^{(i)}$, and let

$$\mathcal{U} = \bigcup_i \mathcal{U}^{(i)} \,. \tag{3.64}$$

Each $\mathcal{U}^{(i)}$ has measure 0, so $\mathcal{U}$ also has measure 0. Let $\boldsymbol{x} \in [0, M]^n \setminus \mathcal{U}$. For each i, the vector $\boldsymbol{x}$ is in the basin of attraction of some $\boldsymbol{a}^{(i)} \in \mathcal{A}_i$. Therefore, there exists some time t_a such that $\left\| \boldsymbol{X}^{(i)}(t_a, \boldsymbol{x}) - \boldsymbol{a}^{(i)} \right\|_2 = \delta$. We refer to such a time as a *crossing time*. Moreover, because $2\delta \leq \delta\left(\boldsymbol{a}^{(i)}\right)$, it follows from (3.60) that $\left\| \boldsymbol{X}^{(i)}(t, \boldsymbol{x}) - \boldsymbol{a}^{(i)} \right\|_2$ is monotonically decreasing in t on some interval $(t_a - \eta, t_a + \eta)$ for all $t \geq t_a$. Therefore, the crossing time t_a is the unique time that satisfies the equality $\left\| \boldsymbol{X}^{(i)}(t_a, \boldsymbol{x}) - \boldsymbol{a}^{(i)} \right\|_2 = \delta$. Because the crossing time is unique, we can define a function $\mathcal{T}^{(i)}(\boldsymbol{x})$ such that $\left\| \boldsymbol{X}^{(i)}\left(\mathcal{T}^{(i)}(\boldsymbol{x}), \boldsymbol{x}\right) - \boldsymbol{a}^{(i)} \right\|_2 = \delta$. This function $\mathcal{T}^{(i)}(\boldsymbol{x})$ gives the unique crossing time for a flow that starts at $\boldsymbol{x}$. gives this unique crossing time.

The local-dynamics function $g^{(i)}$ is continuous, so

$$\mathcal{S}(t, \boldsymbol{x}) = \left\| \boldsymbol{X}^{(i)}(t, \boldsymbol{x}) - \boldsymbol{a}^{(i)} \right\|_2 - \delta \tag{3.65}$$

is continuously differentiable with respect to both t and $\boldsymbol{x}$. Evaluating $\mathcal{S}(t, \boldsymbol{x})$ at $t = \mathcal{T}^{(i)}(\boldsymbol{x})$ yields $\mathcal{S}\left(\mathcal{T}^{(i)}(\boldsymbol{x}), \boldsymbol{x}\right) = 0$. The function $\mathcal{S}(t, \boldsymbol{x})$ is monotonically decreasing in t on some interval $(\mathcal{T}^{(i)}(\boldsymbol{x}) - \eta, \mathcal{T}^{(i)}(\boldsymbol{x}) + \eta)$. Therefore, by the Implicit Function Theorem, $\mathcal{T}^{(i)}$ is a continuous function on each basin of attraction of $g^{(i)}$.

Because $\mathcal{T}^{(i)}(\boldsymbol{x})$ is continuous and each set $\mathcal{F}(\boldsymbol{a}^{(i)}, \boldsymbol{a}^{(j)})$ is compact, there exist times

$$\tau_{ij} = \max_{\boldsymbol{a}^{(i)} \in \mathcal{A}_i, \boldsymbol{a}^{(j)} \in \mathcal{A}_j} \left\{ \mathcal{T}^{(i)}(\boldsymbol{x}) \;\text{ for }\; \boldsymbol{x} \in \mathcal{F}\left(\boldsymbol{a}^{(i)}, \boldsymbol{a}^{(j)}\right) \right\} , \tag{3.66}$$

$$\tau = \max_{\{i,j \;\mid\; \left(H^{(i)}, H^{(j)}\right) \in E\}} \tau_{ij} \,. \tag{3.67}$$

Suppose that an interaction occurs between hosts $H^{(i)}$ and $H^{(j)}$ at time $t_{I,1}$. Let

$$\begin{aligned} \boldsymbol{N}^{(i)}\left(t_{I,1}^{-}\right) &\in B\left(\boldsymbol{a}^{(i)}, \delta\right) , \\ \boldsymbol{N}^{(j)}\left(t_{I,1}^{-}\right) &\in B\left(\boldsymbol{a}^{(j)}, \delta\right) \end{aligned} \tag{3.68}$$

for some $\boldsymbol{a}^{(i)} \in \mathcal{A}_i$ and $\boldsymbol{a}^{(j)} \in \mathcal{A}_j$. We then have

$$\begin{aligned} \boldsymbol{N}^{(i)}\left(t_{I,1}^{+}\right) &\in \mathcal{F}\left(\boldsymbol{a}^{(i)}, \boldsymbol{a}^{(j)}\right) , \\ \boldsymbol{N}^{(j)}\left(t_{I,1}^{+}\right) &\in \mathcal{F}\left(\boldsymbol{a}^{(j)}, \boldsymbol{a}^{(i)}\right) . \end{aligned} \tag{3.69}$$

If the next interaction occurs at time $t_{I,2} > t_{I,1} + \tau$, then

$$\begin{aligned} \boldsymbol{N}^{(i)}\left(t_{I,2}^{-}\right) &\in B\left(\boldsymbol{b}^{(i)}, \delta\right) , \\ \boldsymbol{N}^{(j)}\left(t_{I,2}^{-}\right) &\in B\left(\boldsymbol{b}^{(j)}, \delta\right) \end{aligned} \tag{3.70}$$

for some $\boldsymbol{b}^{(i)} \in \mathcal{A}_i$ and $\boldsymbol{b}^{(j)} \in \mathcal{A}_j$. If no two interactions occur within time τ of each other, then

$$\boldsymbol{N}^{(i)}\left(t_{I}^{-}\right) \in B\left(\boldsymbol{a}^{(i)}, \delta\right) \tag{3.71}$$

for all i, all interaction times t_I, and some $\boldsymbol{a}^{(i)} \in \mathcal{A}_i$. In this case, the effect of each interaction on the basin probability tensor Ψ is described exactly by the operation of the interaction operator

$$\left(\phi\left(\Psi\left(t_{I}^{-}\right)\right)\right)_{b_1,\ldots,b_{|H|}} = \Phi_{b_1,\ldots,b_{|H|},a_1,\ldots,a_{|H|}} \Psi_{a_1,\ldots,a_{|H|}}\left(t_{I}^{-}\right) . \tag{3.72}$$

Let $\mathcal{I}$ be the set of all possible sets $\Omega = \{t_l^*\}_{l=1}^{L}$ of frequency-scaled interaction times in the interval $[0, T^*]$. We select a set Ω of interaction times using a Poisson process on $[0, T^*]$ with rate parameter 1. Let $q : I \to R^+$ be the probability density function for these interactions. Therefore,

$$\Pr\left(\Omega \in \mathcal{J}\right) = \int_{\mathcal{J}} q\left(\Omega'\right) \mathrm{d}\Omega' \tag{3.73}$$

for any $\mathcal{J} \subseteq \mathcal{I}$. We define a counting function

$$\nu(\Omega, t^*) = \left| \left\{ t_l^* \in \Omega \mid t_l^* \leq t^* \right\} \right| \tag{3.74}$$

that tracks the number of interactions that occur in the interval $[0, t^*]$ for the set Ω. The approximate basin probability tensor $\widetilde{\Psi}(t^*)$ from the LFA (see (3.21)) conditioned on Ω is

$$\widetilde{\Psi}^{(\Omega)}(t^*) = \phi^{\nu(\Omega,t^*)}\left(\Psi(0)\right) . \tag{3.75}$$

Therefore,

$$\widetilde{\Psi}(t^*) = \int_{\mathcal{I}} q(\Omega')\widetilde{\Psi}^{(\Omega')}(t^*)\,\mathrm{d}\Omega' . \tag{3.76}$$

Let $\Psi^{(\Omega)}(t^*)$ be the basin probability tensor conditioned on Ω. We then have

$$\Psi(t^*) = \int_{\mathcal{I}} q(\Omega')\Psi^{(\Omega')}(t^*)\,\mathrm{d}\Omega' . \tag{3.77}$$

If $\Omega = \{t_l^*\}_{l=1}^L$ satisfies

$$\min_{l\in\{2,\ldots,L\}}\{t_l^* - t_{l-1}^*\} > \tau^* , \tag{3.78}$$

then

$$\Psi^{(\Omega)}(t^*) = \Phi^{\nu(\Omega,t^*)}\left(\Psi(0)\right) . \tag{3.79}$$

Let $\mathcal{I}_1 \subset \mathcal{I}$ be the set of interaction sets $\Omega = \{t_l^*\}_{l=1}^L$ for which (3.78) holds, and let $\mathcal{I}_2 = \mathcal{I} \setminus \mathcal{I}_1 \subset \mathcal{I}$. Choose dt^* such that $\tau^* < dt^* < 2\tau^*$ and T^*/dt^* is an integer. For sufficiently small τ^*, this is always possible. If two interaction times t_l^* and t_{l-1}^* occur within τ^* of each other, then both interactions occur in an interval $[k\,dt^*, (k-1)dt^*]$ and/or an interval $\left[\left(k+\frac{1}{2}\right)dt^*, \left(k-\frac{1}{2}\right)dt^*\right]$ for an integer k. (These two types of intervals overlap.) All of these intervals have width dt^*. Therefore, the probability that at least two interactions occur in a specific one of these intervals is

$$\begin{aligned}
\Pr\left(\nu\left(\Omega, dt^*\right) \geq 2\right) &= 1 - e^{dt^*} - dt^* e^{dt^*} \\
&= 1 - e^{dt^*}(1 - dt^*) \\
&\leq 1 - (1 + dt^*)(1 - dt^*) \\
&= (dt^*)^2 .
\end{aligned} \tag{3.80}$$

There are $2\frac{T^*}{dt^*} - 1$ such intervals. Therefore, the probability that at least two interactions occur in at least one of these intervals is less than $\left(2\frac{T^*}{dt^*} - 1\right)(dt^*)^2$. This probability equals the probability that a set Ω of interactions satisfies $\Omega \in \mathcal{I}_2$. Consequently,

$$\begin{aligned}
\Pr\left(\Omega \in \mathcal{I}_2\right) &\leq \left(2\frac{T^*}{dt^*} - 1\right)(dt^*)^2 \\
&< 2\frac{T^*}{dt^*}(dt^*)^2 \\
&= 2T^* dt^* \\
&\leq 4T^*\tau^* \\
&= 4T^*\lambda_{\text{tot}}\tau\,.
\end{aligned} \tag{3.81}$$

We choose $\lambda_{\text{tot}} < \frac{\varepsilon}{4T^*\tau}$, and we then have

$$\begin{aligned}
\Psi(t^*) - \widetilde{\Psi}(t^*) &= \int_{\mathcal{I}} q(\Omega')\Psi^{(\Omega')}(t^*)\,\mathrm{d}\Omega' - \int_{\mathcal{I}} q(\Omega')\widetilde{\Psi}^{(\Omega')}(t^*)\,\mathrm{d}\Omega' \\
&= \int_{\mathcal{I}} q(\Omega')\left[\Psi^{(\Omega')}(t^*) - \widetilde{\Psi}^{(\Omega')}(t^*)\right]\mathrm{d}\Omega' \\
&= \int_{\mathcal{I}_1} q(\Omega')\left[\Psi^{(\Omega')}(t^*) - \widetilde{\Psi}^{(\Omega')}(t^*)\right]\mathrm{d}\Omega' + \int_{\mathcal{I}_2} q(\Omega')\left[\Psi^{(\Omega')}(t^*) - \widetilde{\Psi}^{(\Omega')}(t^*)\right]\mathrm{d}\Omega' \\
&= \int_{\mathcal{I}_2} q(\Omega')\left[\Psi^{(\Omega')}(t^*) - \widetilde{\Psi}^{(\Omega')}(t^*)\right]\mathrm{d}\Omega'\,.
\end{aligned} \tag{3.82}$$

Therefore,

$$\left\|\Psi(t^*) - \widetilde{\Psi}(t^*)\right\|_\infty \leq \int_{\mathcal{I}_2} q(\Omega')\left\|\Psi^{(\Omega')}(t^*) - \widetilde{\Psi}^{(\Omega')}(t^*)\right\|_\infty \mathrm{d}\Omega'\,. \tag{3.83}$$

Each entry of $\Psi(t^*)$ and $\widetilde{\Psi}(t^*)$ is a probability and hence is in the interval $[0, 1]$, so we know that $\left\|\Psi(t^*) - \widetilde{\Psi}(t^*)\right\|_\infty \leq 1$. Therefore,

$$\begin{aligned}
\left\|\Psi(t^*) - \widetilde{\Psi}(t^*)\right\|_\infty &\leq \int_{\mathcal{I}_2} q(\Omega')\mathrm{d}\Omega' \\
&= \Pr\left(\Omega \in \mathcal{I}_2\right) \\
&\leq 4T^*\lambda_{\text{tot}}\tau \\
&< \varepsilon\,.
\end{aligned} \tag{3.84}$$

This bound holds for all $t^* \in [0, T^*]$. Because $\varepsilon > 0$ is arbitrary, the basin probability tensor $\Psi(t)$ converges uniformly to $\widetilde{\Psi}(t^*)$ on $[0, T^*]$ as $\lambda_{\text{tot}} \to 0$. □

3.7.2 Proof of High-Frequency Low-Strength Approximation Theorem

In this appendix, we prove the HFLSA Theorem (see Theorem 2).

Theorem 2 (High-Frequency, Low-Strength Approximation Theorem)**.** *Fix the relative interaction-frequency parameters l_{ij}, the product $\lambda_{\text{tot}}\gamma$, and a time T. Let each local-dynamics function $g^{(i)}$ be continuously differentiable and bounded (see Section 3.2.3), and let $\varepsilon \in (0, 1]$ and $\delta > 0$ be arbitrary but fixed. For sufficiently large λ_{tot}, each host's microbiome abundance vector $\boldsymbol{N}^{(i)}(t)$ satisfies*

$$\left\| \boldsymbol{N}^{(i)} - \widetilde{\boldsymbol{N}}^{(i)} \right\|_{L^\infty[0,T]} < \delta \tag{3.22}$$

with probability larger than $1 - \varepsilon$, where

$$\begin{aligned} \frac{d\widetilde{\boldsymbol{N}}^{(i)}}{dt} &= g^{(i)}(\widetilde{\boldsymbol{N}}^{(i)}) + \sum_j \lambda_{ij}\gamma \left(\widetilde{\boldsymbol{N}}^{(j)} - \widetilde{\boldsymbol{N}}^{(i)} \right) , \\ \widetilde{\boldsymbol{N}}^{(i)}(0) &= \boldsymbol{N}^{(i)}(0) . \end{aligned} \tag{3.23}$$

Proof. Each local-dynamics function $g^{(i)}$ is bounded (see Section 3.2.3), so there exists a constant M such that

$$\left\| \boldsymbol{N}^{(i)}(t) \right\|_\infty \leq M \tag{3.85}$$

and each entry of $\boldsymbol{N}^{(i)}(t)$ is nonnegative for each microbiome abundance vector $\boldsymbol{N}^{(i)}(t)$ and all times $t \geq 0$. Each approximate microbiome abundance vector $\widetilde{\boldsymbol{N}}^{(i)}(t)$ also satisfies

$$\left\| \widetilde{\boldsymbol{N}}^{(i)}(t) \right\|_\infty \leq M \tag{3.86}$$

because

$$\begin{aligned} \left(\sum_j \lambda_{ij}\gamma \left(\widetilde{\boldsymbol{N}}^{(j)} - \widetilde{\boldsymbol{N}}^{(i)} \right) \right)_x &\geq 0 \quad \text{if} \quad \widetilde{\boldsymbol{N}}_x^{(i)} = 0 , \\ \left(\sum_j \lambda_{ij}\gamma \left(\widetilde{\boldsymbol{N}}^{(j)} - \widetilde{\boldsymbol{N}}^{(i)} \right) \right)_x &\leq 0 \quad \text{if} \quad \widetilde{\boldsymbol{N}}_x^{(i)} = M \end{aligned} \tag{3.87}$$

for each entry x of $\widetilde{\boldsymbol{N}}^{(i)}(t)$.

We also assume that each local-dynamics function $g^{(i)}$ is continuously differentiable and hence continuous. Each host's microbiome abundance vector $\boldsymbol{N}^{(i)}(t)$ is in the region $[0, M]^n$, which is compact, so there exist constants G and F such that all $\boldsymbol{N}^{(i)}(t)$ satisfy the bounds

$$\begin{aligned} \left\| \frac{\mathrm{d}\boldsymbol{N}^{(i)}}{\mathrm{d}t} \right\|_\infty &= \left\| g^{(i)}\left(\boldsymbol{N}^{(i)}\right) \right\|_\infty \leq G\,, \\ \left\| \frac{\mathrm{d}^2\boldsymbol{N}^{(i)}}{\mathrm{d}t^2} \right\|_\infty &= \left\| Dg^{(i)}\left(\boldsymbol{N}^{(i)}\right) \cdot g^{(i)}\left(\boldsymbol{N}^{(i)}\right) \right\|_\infty \leq 2F\,. \end{aligned} \tag{3.88}$$

For an interaction involving host $H^{(i)}$ that occurs at time t_I, we choose $\boldsymbol{N}^{(i)}\left(t_I\right) = \boldsymbol{N}^{(i)}\left(t_I^+\right)$. Therefore, $\boldsymbol{N}^{(i)}(t)$ is right-continuous at time t_I. It is usually not left-continuous at time t_I, so it is usually not left-differentiable at time t_I.[2] In such situations, the derivatives that we use in (3.88) are right derivatives. By Taylor's theorem,

$$\begin{aligned} \left\| \boldsymbol{N}^{(i)}(t+dt) - \boldsymbol{N}^{(i)}(t) \right\|_\infty &\leq G\,dt\,, \\ \left\| \boldsymbol{N}^{(i)}(t+dt) - \boldsymbol{N}^{(i)}(t) - g^{(i)}\left(N^i(t)\right) dt \right\|_\infty &\leq F\,dt^2 \end{aligned} \tag{3.89}$$

for any interval $(t, t+dt]$ in which there are no interactions. Each local-dynamics function $g^{(i)}$ is also Lipschitz continuous. Therefore, there exists a constant C such that

$$\left\| g^{(i)}\left(\boldsymbol{x}\right) - g^{(i)}\left(\boldsymbol{y}\right) \right\|_\infty \leq C \left\| \boldsymbol{x} - \boldsymbol{y} \right\|_\infty \tag{3.90}$$

for each $g^{(i)}$ and all $\boldsymbol{x}, \boldsymbol{y} \in [0, M]^n$.

Let $\widetilde{G} = G + M\lambda_{\mathrm{tot}}\gamma$, which is a constant because $\lambda_{\mathrm{tot}}\gamma$ is fixed. For all host microbiome abundance vectors $\widetilde{\boldsymbol{N}}^{(i)}(t)$, we have

$$\begin{aligned} \left\| \frac{\mathrm{d}\widetilde{\boldsymbol{N}}^{(i)}}{\mathrm{d}t} \right\|_\infty &= \left\| g^{(i)}\left(\widetilde{\boldsymbol{N}}^{(i)}\right) + \sum_j \lambda_{ij}\gamma \left(\widetilde{\boldsymbol{N}}^{(j)} - \widetilde{\boldsymbol{N}}^{(i)}\right) \right\|_\infty \\ &\leq G + \sum_j \lambda_{ij}\gamma \left\| \widetilde{\boldsymbol{N}}^{(j)} - \widetilde{\boldsymbol{N}}^{(i)} \right\|_\infty \leq G + M\lambda_{\mathrm{tot}}\gamma = \widetilde{G}\,. \end{aligned} \tag{3.91}$$

[2]The only situation where $\boldsymbol{N}^{(i)}(t)$ is left-continuous at time t_I occurs when the host $H^{(j)}$ with which $H^{(i)}$ interacts has a microbiome vector $\boldsymbol{N}^{(j)}(t_I^-) = \boldsymbol{N}^{(i)}(t_I^-)$.

Let $\widetilde{F} = F + G\lambda_{\text{tot}}\gamma + M\lambda_{\text{tot}}^2\gamma^2$. For all $\widetilde{\boldsymbol{N}}^{(i)}(t)$, we have

$$\begin{aligned}\left\|\frac{\mathrm{d}^2\widetilde{\boldsymbol{N}}^{(i)}}{\mathrm{d}t^2}\right\|_\infty &= \left\| \begin{array}{l} Dg^{(i)}\left(\widetilde{\boldsymbol{N}}^{(i)}\right)\cdot g^{(i)}\left(\widetilde{\boldsymbol{N}}^{(i)}\right) \\ \quad + \sum_j \lambda_{ij}\gamma\left[g^{(j)}\left(\widetilde{\boldsymbol{N}}^{(j)}\right) + \sum_k \lambda_{jk}\gamma\left(\widetilde{\boldsymbol{N}}^{(k)} - \widetilde{\boldsymbol{N}}^{(j)}\right)\right] \\ \quad + \sum_j \lambda_{ij}\gamma\left[g^{(i)}\left(\widetilde{\boldsymbol{N}}^{(i)}\right) + \sum_k \lambda_{ik}\gamma\left(\widetilde{\boldsymbol{N}}^{(k)} - \widetilde{\boldsymbol{N}}^{(i)}\right)\right] \end{array} \right\|_\infty \\ &\leq 2F + G\gamma\sum_j \lambda_{ij} + M\gamma^2\sum_j\sum_k \lambda_{ij}\lambda_{jk} + G\gamma\sum_j \lambda_{ij} + M\gamma^2\sum_j\sum_k \lambda_{ij}\lambda_{ik} \\ &\leq 2F + 2G\lambda_{\text{tot}}\gamma + 2M\lambda_{\text{tot}}^2\gamma^2 = 2\widetilde{F}\,.\end{aligned} \tag{3.92}$$

By Taylor's theorem,

$$\left\| \widetilde{\boldsymbol{N}}^{(i)}(t+dt) - \widetilde{\boldsymbol{N}}^{(i)}(t) \right\|_\infty \leq \widetilde{G}\,dt\,, \tag{3.93}$$

$$\left\| \widetilde{\boldsymbol{N}}^{(i)}(t+dt) - \widetilde{\boldsymbol{N}}^{(i)}(t) - \left[g^{(i)}\left(\widetilde{\boldsymbol{N}}^{(i)}(t)\right) + \sum_j \lambda_{ij}\gamma\left(\widetilde{\boldsymbol{N}}^{(j)}(t) - \widetilde{\boldsymbol{N}}^{(i)}(t)\right)\right] dt \right\|_\infty \leq \widetilde{F}\,dt^2$$

for any times t and $t + dt$.

We phrased Theorem 2 in terms of finding a sufficiently large total-interaction-frequency parameter λ_{tot} so that

$$\left\|\boldsymbol{N}^{(i)} - \widetilde{\boldsymbol{N}}^{(i)}\right\|_{L^\infty[0,T]} < \delta \tag{3.94}$$

with probability larger than $1 - \varepsilon$. In Theorem 2, we assume that the product $\lambda_{\text{tot}}\gamma$ is fixed, so finding a sufficiently large λ_{tot} is equivalent to finding a sufficiently small γ. We define the error term

$$E^{(i)}(t) = \boldsymbol{N}^{(i)}(t) - \widetilde{\boldsymbol{N}}^{(i)}(t)\,. \tag{3.95}$$

We will show that one can bound each $\left\|E^{(i)}(t)\right\|_\infty$ with probability larger than $1 - \varepsilon$ by a term that involves γ. For sufficiently small γ, we will show that each

$$\left\|E^{(i)}(t)\right\|_\infty \leq \delta \tag{3.96}$$

for all $t \in [0, T]$ with probability larger than $1 - \varepsilon$.

Fix a dt such that $dt^4 \leq \gamma \leq 4\,dt^4 < 1$ and T/dt is an integer. This is always possible for sufficiently small γ. Let $t_k = k\,dt$. There are only finitely many t_k in the interval $[0, T]$. The probability that an interaction occurs precisely at any of these t_k is 0. Therefore, for the remainder of this proof, we only consider interactions that occur at times $t \neq t_k$ for any k. Under this assumption,

$$\boldsymbol{N}^{(i)}(t_k^-) = \boldsymbol{N}^{(i)}(t_k) \tag{3.97}$$

for all $t_k \in [0, T]$.

We now consider how a microbiome abundance vector $\boldsymbol{N}^{(i)}(t)$ changes over an interval $[t_k, t_{k+1}]$. Let L_k be the number of interactions that occur in (t_k, t_{k+1}). (This interval is open because no interactions occur at any of the t_k.) We denote the associated ordered set of interactions by $\{t_{k,l}\}_{l=1}^{L_k}$. Additionally, we let $t_{k,0} = t_k$ and $t_{k,L_k+1} = t_{k+1}$, and we define $dt_{k,l} = t_{k,l} - t_{k,l-1}$. For $l \in \{1, \ldots, L_k + 1\}$, let

$$A_{k,l}^{(i)} = \boldsymbol{N}^{(i)}\left(t_{k,l}^-\right) - \boldsymbol{N}^{(i)}(t_k)\,. \tag{3.98}$$

For $l \in \{1, \ldots, L_k\}$, let

$$J_{k,l}^{(i)} = \boldsymbol{N}^{(i)}\left(t_{k,l}^+\right) - \boldsymbol{N}^{(i)}\left(t_{k,l}^-\right)\,. \tag{3.99}$$

The difference $J_{k,l}^{(i)}$ indicates the change in $\boldsymbol{N}^{(i)}(t)$ after an interaction at time $t_{k,l}$. If this interaction does not involve host $H^{(i)}$, then $J_{k,l}^{(i)} = \boldsymbol{0}$. Otherwise, for an interaction between hosts $H^{(i)}$ and $H^{(j)}$, we have

$$J_{k,l}^{(i)} = \gamma\left(\boldsymbol{N}^{(j)}\left(t_{k,l}^-\right) - \boldsymbol{N}^{(i)}\left(t_{k,l}^-\right)\right)\,. \tag{3.100}$$

In either case, $\left\|J_{k,l}^{(i)}\right\|_\infty \leq M\gamma$. For $l \geq 2$, we decompose the difference $A_{k,l}^{(i)}$ by writing

$$\begin{aligned} A_{k,l}^{(i)} &= \boldsymbol{N}^{(i)}\left(t_{k,l}^-\right) - \boldsymbol{N}^{(i)}(t_{k,l-1}^+) + \boldsymbol{N}^{(i)}(t_{k,l-1}^+) - \boldsymbol{N}^{(i)}(t_{k,l-1}^-) \\ &\quad + \boldsymbol{N}^{(i)}(t_{k,l-1}^-) - \boldsymbol{N}^{(i)}(t_k) \\ &= \boldsymbol{N}^{(i)}\left(t_{k,l}^-\right) - \boldsymbol{N}^{(i)}(t_{k,l-1}^+) + J_{k,l-1}^{(i)} + A_{k,l-1}^{(i)}\,. \end{aligned} \tag{3.101}$$

Therefore,

$$\begin{aligned}\left\|A_{k,l}^{(i)}\right\|_\infty &\le \left\|\boldsymbol{N}^{(i)}\left(t_{k,l}^{-}\right)-\boldsymbol{N}^{(i)}(t_{k,l-1}^{+})\right\|_\infty+\left\|J_{k,l-1}^{(i)}\right\|_\infty+\left\|A_{k,l-1}^{(i)}\right\|_\infty \\ &\le \left\|A_{k,l-1}^{(i)}\right\|_\infty+G\,dt_{k,l}+M\gamma\,.\end{aligned} \tag{3.102}$$

For $l=1$, we have

$$\left\|A_{k,1}^{(i)}\right\|_\infty \le \left\|\boldsymbol{N}^{(i)}(t_{k,1}^{-})-\boldsymbol{N}^{(i)}(t_k)\right\|_\infty \le G\,dt_{k,1}\,. \tag{3.103}$$

Therefore,

$$\left\|A_{k,l}^{(i)}\right\|_\infty \le \sum_{l'=1}^{l} G\,dt_{k,l'}+(l-1)M\gamma \le \sum_{l'=1}^{L_k+1} G\,dt_{k,l'}+L_kM\gamma = G\,dt+L_kM\gamma\,. \tag{3.104}$$

The error between the actual microbiome abundance vector $\boldsymbol{N}^{(i)}(t_k)$ and the approximate microbiome abundance vector $\widetilde{\boldsymbol{N}}^{(i)}(t_k)$ is

$$E_k^{(i)} = \boldsymbol{N}^{(i)}(t_k)-\widetilde{\boldsymbol{N}}^{(i)}(t_k)\,. \tag{3.105}$$

Consider the difference

$$\begin{aligned}E_{k+1}^{(i)}-E_k^{(i)} &= \left[\boldsymbol{N}^{(i)}(t_{k+1})-\widetilde{\boldsymbol{N}}^{(i)}(t_{k+1})\right]-\left[\boldsymbol{N}^{(i)}(t_k)-\widetilde{\boldsymbol{N}}^{(i)}(t_k)\right] \\ &= \left[\boldsymbol{N}^{(i)}(t_{k+1})-\boldsymbol{N}^{(i)}(t_k)\right]-\left[\widetilde{\boldsymbol{N}}^{(i)}(t_{k+1})-\widetilde{\boldsymbol{N}}^{(i)}(t_k)\right]\,.\end{aligned} \tag{3.106}$$

We have

$$\widetilde{\boldsymbol{N}}^{(i)}(t_{k+1})-\widetilde{\boldsymbol{N}}^{(i)}(t_k) = \left[g^{(i)}\left(\widetilde{\boldsymbol{N}}^{(i)}(t_k)\right)+\sum_j \lambda_{ij}\gamma\left(\widetilde{\boldsymbol{N}}^{(j)}(t_k)-\widetilde{\boldsymbol{N}}^{(i)}(t_k)\right)\right]dt+\eta_k^{(i),\,\mathrm{approx}}, \tag{3.107}$$

where

$$\left\|\eta_k^{(i),\,\mathrm{approx}}\right\|_\infty \le \widetilde{F}\,dt^2\,. \tag{3.108}$$

The change of the microbiome abundance vector $\boldsymbol{N}^{(i)}(t)$ over the interval $[t_k, t_{k+1}]$ is

$$\begin{aligned}
\boldsymbol{N}^{(i)}(t_{k+1}) - \boldsymbol{N}^{(i)}(t_k) &= \sum_{l=1}^{L_k+1} \left[\boldsymbol{N}^{(i)}\left(t_{k,l}^-\right) - \boldsymbol{N}^{(i)}\left(t_{k,l-1}^-\right)\right] \\
&= \sum_{l=1}^{L_k+1} \left[\boldsymbol{N}^{(i)}\left(t_{k,l}^-\right) - \boldsymbol{N}^{(i)}\left(t_{k,l-1}^+\right)\right] + \sum_{l=2}^{L_k+1} \left[\boldsymbol{N}^{(i)}\left(t_{k,l-1}^+\right) - \boldsymbol{N}^{(i)}\left(t_{k,l-1}^-\right)\right] \\
&= \sum_{l=1}^{L_k+1} \left[g^{(i)}\left(\boldsymbol{N}^{(i)}\left(t_k\right)\right) dt_{k,l} + \eta_{k,l}^{(i),\,\mathrm{local}}\right] + \sum_{l=1}^{L_k} J_{k,l}^{(i)} \\
&= g^{(i)}\left(\boldsymbol{N}^{(i)}\left(t_k\right)\right) dt + \sum_{l=1}^{L_k+1} \left[\eta_{k,l}^{(i),\,\mathrm{local}}\right] + \sum_{l=1}^{L_k} J_{k,l}^{(i)}\,,
\end{aligned} \tag{3.109}$$

where

$$\begin{aligned}
\eta_{k,l}^{(i),\,\mathrm{local}} = &\left[\boldsymbol{N}^{(i)}\left(t_{k,l}^-\right) - \boldsymbol{N}^{(i)}\left(t_{k,l-1}^+\right) - g^{(i)}\left(\boldsymbol{N}^{(i)}\left(t_{k,l}^-\right)\right) dt_{k,l}\right] \\
&+ \left[g^{(i)}\left(\boldsymbol{N}^{(i)}\left(t_{k,l}^-\right)\right) dt_{k,l} - g^{(i)}\left(\boldsymbol{N}^{(i)}\left(t_k\right)\right) dt_{k,l}\right]\,.
\end{aligned} \tag{3.110}$$

Therefore,

$$\begin{aligned}
\left\|\eta_{k,l}^{(i),\,\mathrm{local}}\right\|_\infty &\le F\, dt_{k,l}^2 + C \left\|\boldsymbol{N}^{(i)}\left(t_{k,l}^-\right) - \boldsymbol{N}^{(i)}\left(t_k\right)\right\|_\infty dt_{k,l} \\
&= F\, dt_{k,l}^2 + C \left\|A_{k,l}^{(i)}\right\|_\infty dt_{k,l} \\
&\le F\, dt_{k,l}^2 + CG\, dt\, dt_{k,l} + CL_kM\gamma\, dt_{k,l} \\
&\le \left(F\, dt + CG\, dt + CL_kM\gamma\right) dt_{k,l}\,.
\end{aligned} \tag{3.111}$$

We now consider

$$\sum_{l=1}^{L_k} J_{k,l}^{(i)}\,, \tag{3.112}$$

which is the sum of the changes in microbiome abundance vector $\boldsymbol{N}^{(i)}(t)$ due to the interactions that host $H^{(i)}$ has with other hosts. Let

$$\widetilde{J}_{k,l}^{(i)} = \begin{cases} \boldsymbol{0}\,, & \text{the interaction at time } t_{k,l} \text{ does not involve } H^{(i)} \\ \gamma\left(\boldsymbol{N}^{(j)}(t_k) - \boldsymbol{N}^{(i)}(t_k)\right)\,, & \text{the interaction at time } t_{k,l} \text{ is between } H^{(i)} \text{ and } H^{(j)}\,. \end{cases} \tag{3.113}$$

We then have

$$\sum_{l=1}^{L_k} J_{k,l}^{(i)} = \sum_{l=1}^{L_k} \left(\widetilde{J}_{k,l}^{(i)} + \eta_{k,l}^{(i),\,\text{exchange}} \right) , \tag{3.114}$$

where

$$\eta_{k,l}^{(i),\,\text{exchange}} = J_{k,l}^{(i)} - \widetilde{J}_{k,l}^{(i)} . \tag{3.115}$$

If the interaction at time $t_{k,l}$ does not involve host $H^{(i)}$, then $\eta_{k,l}^{(i),\,\text{exchange}} = \mathbf{0}$. Otherwise, for an interaction at time $t_{k,l}$ between hosts $H^{(i)}$ and $H^{(j)}$, we have

$$\begin{aligned} \eta_{k,l}^{(i),\,\text{exchange}} &= \gamma \left(\boldsymbol{N}^{(j)}(t_k) - \boldsymbol{N}^{(i)}(t_k) \right) - \gamma \left(\boldsymbol{N}^{(j)} \left(t_{k,l}^- \right) - \boldsymbol{N}^{(i)} \left(t_{k,l}^- \right) \right) \\ &= \gamma \left[\left(\boldsymbol{N}^{(j)}(t_k) - \boldsymbol{N}^{(j)} \left(t_{k,l}^- \right) \right) - \left(\boldsymbol{N}^{(i)}(t_k) - \boldsymbol{N}^{(i)} \left(t_{k,l}^- \right) \right) \right] . \end{aligned} \tag{3.116}$$

In either case,

$$\begin{aligned} \left\| \eta_{k,l}^{(i),\,\text{exchange}} \right\|_\infty &\leq \gamma \left(\left\| \boldsymbol{N}^{(j)}(t_k) - \boldsymbol{N}^{(j)} \left(t_{k,l}^- \right) \right\|_\infty + \left\| \boldsymbol{N}^{(i)}(t_k) - \boldsymbol{N}^{(i)} \left(t_{k,l}^- \right) \right\|_\infty \right) \\ &= \gamma \left(\left\| A_{k,l}^{(j)} \right\|_\infty + \left\| A_{k,l}^{(i)} \right\|_\infty \right) \\ &\leq 2G\gamma \, dt + 2L_k M \gamma^2 . \end{aligned} \tag{3.117}$$

We seek to bound

$$\left\| \sum_{l=1}^{L_k} \widetilde{J}_{k,l}^{(i)} - \mathbb{E} \left[\sum_{l=1}^{L_k} \widetilde{J}_{k,l}^{(i)} \right] \right\|_\infty . \tag{3.118}$$

There are two sources of stochasticity in the sum

$$\sum_{l=1}^{L_k} \widetilde{J}_{k,l}^{(i)} . \tag{3.119}$$

The number L_k of interactions follows a Poisson distribution with mean $\lambda_{\text{tot}} dt$, and the term $\widetilde{J}_{k,l}^{(i)}$ depends on which pair of hosts interacts at time $t_{k,l}$. Because the sum (3.118) is stochastic, we are only able to bound it with high probability.

The quantity $\widetilde{J}_{k,l}^{(i)}$ is vector-valued, and we denote an entry x of it by $\left(\widetilde{J}_{k,l}^{(i)} \right)_x$. The sum (3.119) is also vector-valued, and we denote an entry x of it by $\left(\sum_{l=1}^{L} \widetilde{J}_{k,l}^{(i)} \right)_x$. We now

calculate the expectation and the variance of each entry of (3.119). The interaction at time $t_{k,l}$ is between hosts $H^{(i)}$ and $H^{(j)}$ with probability $\lambda_{ij}/\lambda_{\text{tot}}$. Therefore,

$$\begin{aligned}
\mathbb{E}\left[L_k\right] &= \lambda_{\text{tot}} dt\,, \\
\text{Var}\left[L_k\right] &= \lambda_{\text{tot}} dt\,, \\
\mathbb{E}\left[\left(\widetilde{J}_{k,l}^{(i)}\right)_x\right] &= \sum_j \frac{\lambda_{ij}}{\lambda_{\text{tot}}} \gamma \left(\boldsymbol{N}^{(j)}(t_k) - \boldsymbol{N}^{(i)}(t_k)\right)_x\,, \\
\text{Var}\left[\left(\widetilde{J}_{k,l}^{(i)}\right)_x\right] &= \mathbb{E}\left[\left(\widetilde{J}_{k,l}^{(i)}\right)_x^2\right] - \left(\mathbb{E}\left[\left(\widetilde{J}_{k,l}^{(i)}\right)_x\right]\right)^2.
\end{aligned} \tag{3.120}$$

The expectation of each entry of the sum (3.119) is

$$\begin{aligned}
\mathbb{E}\left[\left(\sum_{l=1}^{L_k} \widetilde{J}_{k,l}^{(i)}\right)_x\right] &= \mathbb{E}\left[\ \mathbb{E}\left[\left(\sum_{l=1}^{L_k} \widetilde{J}_{k,l}^{(i)}\right)_x \,\middle|\, L_k\right]\ \right] \\
&= \mathbb{E}\left[L_k \sum_j \frac{\lambda_{ij}}{\lambda_{\text{tot}}} \gamma \left(\boldsymbol{N}^{(j)}(t_k) - \boldsymbol{N}^{(i)}(t_k)\right)_x\right] \\
&= \lambda_{\text{tot}} dt \sum_j \frac{\lambda_{ij}}{\lambda_{\text{tot}}} \gamma \left(\boldsymbol{N}^{(j)}(t_k) - \boldsymbol{N}^{(i)}(t_k)\right)_x \\
&= \sum_j \lambda_{ij} \gamma \left(\boldsymbol{N}^{(j)}(t_k) - \boldsymbol{N}^{(i)}(t_k)\right)_x dt\,.
\end{aligned} \tag{3.121}$$

Therefore,

$$\mathbb{E}\left[\sum_{l=1}^{L_k} \widetilde{J}_{k,l}^{(i)}\right] = \sum_j \lambda_{ij} \gamma \left(\boldsymbol{N}^{(j)}(t_k) - \boldsymbol{N}^{(i)}(t_k)\right) dt\,. \tag{3.122}$$

Applying the law of total variance, the variance of each entry of the sum (3.119) is

$$\operatorname{Var}\left[\left(\sum_{l=1}^{L_k}\widetilde{J}_{k,l}^{(i)}\right)_x\right] = \mathbb{E}\left[\operatorname{Var}\left[\left(\sum_{l=1}^{L_k}\widetilde{J}_{k,l}^{(i)}\right)_x \,\middle|\, L_k\right]\right] + \operatorname{Var}\left[\mathbb{E}\left[\left(\sum_{l=1}^{L_k}\widetilde{J}_{k,l}^{(i)}\right)_x \,\middle|\, L_k\right]\right] \tag{3.123}$$

$$\begin{aligned}
&= \mathbb{E}\left[L_k \operatorname{Var}\left[\left(\widetilde{J}_{k,l}^{(i)}\right)_x\right]\right] + \operatorname{Var}\left[L_k \mathbb{E}\left[\left(\widetilde{J}_{k,l}^{(i)}\right)_x\right]\right] \\
&= \lambda_{\mathrm{tot}} dt \operatorname{Var}\left[\left(\widetilde{J}_{k,l}^{(i)}\right)_x\right] + \lambda_{\mathrm{tot}} dt \left(\mathbb{E}\left[\left(\widetilde{J}_{k,l}^{(i)}\right)_x\right]\right)^2 \\
&= \lambda_{\mathrm{tot}} dt\, \mathbb{E}\left[\left(\widetilde{J}_{k,l}^{(i)}\right)_x^2\right] \\
&= \lambda_{\mathrm{tot}} dt \sum_j \frac{\lambda_{ij}}{\lambda_{\mathrm{tot}}} \gamma^2 \left(\boldsymbol{N}^{(j)}(t_k) - \boldsymbol{N}^{(i)}(t_k)\right)_x^2 \\
&= \sum_j \lambda_{ij} \gamma^2 \left(\boldsymbol{N}^{(j)}(t_k) - \boldsymbol{N}^{(i)}(t_k)\right)_x^2 dt\,.
\end{aligned}$$

These variances satisfy the bound

$$\begin{aligned}
\operatorname{Var}\left[\sum_{l=1}^{L_k}\left(\widetilde{J}_{k,l}^{(i)}\right)_x\right] &\le \sum_j \lambda_{ij}\gamma^2 \left(\boldsymbol{N}^{(j)}(t_k) - \boldsymbol{N}^{(i)}(t_k)\right)_x^2 dt \\
&\le M^2 \gamma^2\, dt \sum_j \lambda_{ij} \le M^2 \lambda_{\mathrm{tot}} \gamma^2\, dt\,.
\end{aligned} \tag{3.124}$$

The error due to the stochasticity of the interactions is

$$\eta_k^{(i),\,\mathrm{exchange}} = \sum_{l=1}^{L_k} \widetilde{J}_{k,l}^{(i)} - \mathbb{E}\left[\sum_{l=1}^{L_k} \widetilde{J}_{k,l}^{(i)}\right]. \tag{3.125}$$

Therefore,

$$\sum_{l=1}^{L_k} \widetilde{J}_{k,l}^{(i)} = \eta_k^{(i),\,\mathrm{exchange}} + \mathbb{E}\left[\sum_{l=1}^{L_k} \widetilde{J}_{k,l}^{(i)}\right] = \eta_k^{(i),\,\mathrm{exchange}} + \sum_j \lambda_{ij}\gamma\left(\boldsymbol{N}^{(j)}(t_k) - \boldsymbol{N}^{(i)}(t_k)\right) dt\,. \tag{3.126}$$

The error (3.125) satisfies

$$\left\|\eta_k^{(i),\,\mathrm{exchange}}\right\|_\infty = \max_x \left\{ \left\| \sum_{l=1}^{L_k}\left(\widetilde{J}_{k,l}^{(i)}\right)_x - \mathbb{E}\left[\sum_{l=1}^{L_k}\left(\widetilde{J}_{k,l}^{(i)}\right)_x\right] \right\| \right\}. \tag{3.127}$$

Let $\kappa = \dim\left(\boldsymbol{N}^{(i)}\right)$ and $\alpha = \sqrt{\frac{3\kappa T}{\varepsilon\, dt}}$. By Chebyshev's inequality,

$$\Pr\left(\left\|\sum_{l=1}^{L_k}\left(\widetilde{J}_{k,l}^{(i)}\right)_x - \mathbb{E}\left[\sum_{l=1}^{L_k}\left(\widetilde{J}_{k,l}^{(i)}\right)_x\right]\right\| \geq \alpha\sqrt{\operatorname{Var}\left[\sum_{l=1}^{L_k}\left(\widetilde{J}_{k,l}^{(i)}\right)_x\right]}\right) \leq \frac{1}{\alpha^2}\,, \tag{3.128}$$

$$\Pr\left(\left\|\sum_{l=1}^{L_k}\left(\widetilde{J}_{k,l}^{(i)}\right)_x - \mathbb{E}\left[\sum_{l=1}^{L_k}\left(\widetilde{J}_{k,l}^{(i)}\right)_x\right]\right\| \geq \alpha M\gamma\sqrt{\lambda_{\text{tot}}dt}\right) \leq \frac{1}{\alpha^2}\,,$$

$$\Pr\left(\left\|\sum_{l=1}^{L_k}\left(\widetilde{J}_{k,l}^{(i)}\right)_x - \mathbb{E}\left[\sum_{l=1}^{L_k}\left(\widetilde{J}_{k,l}^{(i)}\right)_x\right]\right\| \geq M\gamma\sqrt{\frac{3\kappa T\lambda_{\text{tot}}}{\varepsilon}}\right) \leq \frac{\varepsilon\, dt}{3\kappa T}\,.$$

Therefore,

$$\Pr\left(\left\|\eta_k^{(i),\,\text{exchange}}\right\|_\infty < M\gamma\sqrt{\frac{3\kappa T\lambda_{\text{tot}}}{\varepsilon}}\right) > 1 - \frac{\varepsilon\, dt}{3T}\,. \tag{3.129}$$

Inserting (3.126) into (3.114) yields

$$\begin{aligned}\sum_{l=1}^{L_k} J_{k,l}^{(i)} &= \sum_{l=1}^{L_k}\left(\widetilde{J}_{k,l}^{(i)} + \eta_{k,l}^{(i),\,\text{exchange}}\right) \\ &= \sum_j \lambda_{ij}\gamma\left(\boldsymbol{N}^{(j)}(t_k) - \boldsymbol{N}^{(i)}(t_k)\right)dt + \sum_{l=1}^{L_k}\eta_{k,l}^{(i),\,\text{exchange}} + \eta_k^{(i),\,\text{exchange}}\,.\end{aligned} \tag{3.130}$$

Inserting (3.107), (3.109), and (3.130) into (3.106) yields

$$\begin{aligned}E_{k+1}^{(i)} - E_k^{(i)} &= \left[\boldsymbol{N}^{(i)}(t_{k+1}) - \boldsymbol{N}^{(i)}(t_k)\right] - \left[\widetilde{\boldsymbol{N}}^{(i)}(t_{k+1}) - \widetilde{\boldsymbol{N}}^{(i)}(t_k)\right] \\ &= g^{(i)}\left(\boldsymbol{N}^{(i)}\left(t_k\right)\right)dt + \sum_{l=1}^{L_k+1}\eta_{k,l}^{(i),\,\text{local}} \\ &\quad + \sum_j \lambda_{ij}\gamma\left(\boldsymbol{N}^{(j)}(t_k) - \boldsymbol{N}^{(i)}(t_k)\right)dt + \sum_{l=1}^{L_k}\eta_{k,l}^{(i),\,\text{exchange}} + \eta_k^{(i),\,\text{exchange}} \\ &\quad - \left[g^{(i)}\left(\widetilde{\boldsymbol{N}}^{(i)}(t_k)\right) + \sum_j \lambda_{ij}\gamma\left(\widetilde{\boldsymbol{N}}^{(j)}(t_k) - \widetilde{\boldsymbol{N}}^{(i)}(t_k)\right)\right]dt - \eta_k^{(i),\,\text{approx}} \\ &= g^{(i)}\left(\boldsymbol{N}^{(i)}\left(t_k\right)\right)dt - g^{(i)}\left(\widetilde{\boldsymbol{N}}^{(i)}(t_k)\right)dt \\ &\quad + \sum_j \lambda_{ij}\gamma\left(\boldsymbol{N}^{(j)}(t_k) - \boldsymbol{N}^{(i)}(t_k)\right)dt - \sum_j \lambda_{ij}\gamma\left(\widetilde{\boldsymbol{N}}^{(j)}(t_k) - \widetilde{\boldsymbol{N}}^{(i)}(t_k)\right)dt \\ &\quad + \sum_{l=1}^{L_k+1}\eta_{k,l}^{(i),\,\text{local}} + \sum_{l=1}^{L_k}\eta_{k,l}^{(i),\,\text{exchange}} + \eta_k^{(i),\,\text{exchange}} - \eta_k^{(i),\,\text{approx}}\,.\end{aligned} \tag{3.131}$$

Therefore,

$$\begin{aligned}\left\|E_{k+1}^{(i)}\right\|_\infty \leq \left\|E_k^{(i)}\right\|_\infty &+ \left\|g^{(i)}\left(\boldsymbol{N}^{(i)}\left(t_k\right)\right) - g^{(i)}\left(\widetilde{\boldsymbol{N}}^{(i)}(t_k)\right)\right\|_\infty dt \\ &+ \sum_j \lambda_{ij}\gamma\left(\left\|\boldsymbol{N}^{(j)}(t_k) - \widetilde{\boldsymbol{N}}^{(j)}(t_k)\right\|_\infty + \left\|\boldsymbol{N}^{(i)}(t_k) - \widetilde{\boldsymbol{N}}^{(i)}(t_k)\right\|_\infty\right) dt \\ &+ \sum_{l=1}^{L_k+1}\left\|\eta_{k,l}^{(i),\,\mathrm{local}}\right\|_\infty + \sum_{l=1}^{L_k}\left\|\eta_{k,l}^{(i),\,\mathrm{exchange}}\right\|_\infty + \left\|\eta_k^{(i),\,\mathrm{exchange}}\right\|_\infty + \left\|\eta_k^{(i),\,\mathrm{approx}}\right\|_\infty .\end{aligned} \tag{3.132}$$

The inequality (3.129) gives a bound for $\left\|\eta_k^{(i),\,\mathrm{exchange}}\right\|_\infty$ that holds with probability larger than $1-\frac{\varepsilon\,dt}{3T}$. The inequalities (3.111) and (3.117) give bounds for the error terms $\left\|\eta_{k,l}^{(i),\,\mathrm{local}}\right\|_\infty$ and $\left\|\eta_{k,l}^{(i),\,\mathrm{exchange}}\right\|_\infty$, respectively. The latter two bounds depend on the number L_k of interactions in the interval (t_k, t_{k+1}). As we discussed above, L_k is stochastic and is distributed as a Poisson random variable with mean $\lambda_{\mathrm{tot}}dt$. Therefore, we give bounds for L_k that hold with high probability. To obtain these bounds, we first define

$$\beta = \sqrt{\frac{3T}{\varepsilon\,dt}}\,. \tag{3.133}$$

By Chebyshev's inequality,

$$\begin{aligned}\Pr\left(L_k \geq \lambda_{\mathrm{tot}}dt + \beta\sqrt{\lambda_{\mathrm{tot}}dt}\right) &\leq \frac{1}{\beta^2}\,, \\ \Pr\left(L_k \geq \lambda_{\mathrm{tot}}dt + \sqrt{\frac{3T\lambda_{\mathrm{tot}}}{\varepsilon}}\right) &\leq \frac{\varepsilon\,dt}{3T}\,.\end{aligned} \tag{3.134}$$

Therefore, with probability at least $1-\frac{\varepsilon\,dt}{3T}$, we have the bounds

$$L_k \leq \lambda_{\mathrm{tot}}dt + \sqrt{\frac{3T\lambda_{\mathrm{tot}}}{\varepsilon}}\,. \tag{3.135}$$

and

$$\begin{aligned}L_k\gamma &\leq \lambda_{\mathrm{tot}}\gamma\,dt + \sqrt{\gamma}\sqrt{\frac{3T\lambda_{\mathrm{tot}}\gamma}{\varepsilon}} \\ &\leq \lambda_{\mathrm{tot}}\gamma\,dt + 2dt^2\sqrt{\frac{3T\lambda_{\mathrm{tot}}\gamma}{\varepsilon}} \\ &\leq \left(\lambda_{\mathrm{tot}}\gamma + 2\sqrt{\frac{3T\lambda_{\mathrm{tot}}\gamma}{\varepsilon}}\right) dt\,,\end{aligned} \tag{3.136}$$

where we use the inequalities $\gamma < 4dt^4$ and $dt < 1$. With probability at least $1-\left(\frac{\varepsilon\, dt}{3T} + \frac{\varepsilon\, dt}{3T}\right) = 1 - \frac{2\varepsilon\, dt}{3T}$, the bounds in (3.129) and (3.136) both hold, yielding

$$\begin{aligned}
\left\|E_{k+1}^{(i)}\right\|_\infty &\leq \left\|E_k^{(i)}\right\|_\infty + C\left\|E_k^{(i)}\right\|_\infty dt + \left(\left\|E_k^{(j)}\right\|_\infty + \left\|E_k^{(i)}\right\|_\infty\right)\lambda_{\text{tot}}\gamma\, dt \\
&\quad + \sum_{l=1}^{L_k+1}\left(F\, dt + CG\, dt + CL_kM\gamma\right) dt_{k,l} \\
&\quad + \sum_{l=1}^{L_k}\left(2G\gamma\, dt + 2L_kM\gamma^2\right) + M\gamma\sqrt{\frac{3\kappa T\lambda_{\text{tot}}}{\varepsilon}} + \widetilde{F}\, dt^2 \\
&\leq \left(1 + (C + \lambda_{\text{tot}}\gamma)\, dt\right)\left\|E_k^{(i)}\right\|_\infty + \lambda_{\text{tot}}\gamma\, dt\left\|E_k^{(j)}\right\|_\infty \\
&\quad + (F + CG)dt^2 + CL_kM\gamma\, dt + 2GL_k\gamma\, dt \\
&\quad + 2L_k^2M\gamma^2 + M\sqrt{\gamma}\sqrt{\frac{3\kappa T\lambda_{\text{tot}}\gamma}{\varepsilon}} + \widetilde{F}\, dt^2 \\
&\leq \left(1 + (C + \lambda_{\text{tot}}\gamma)\, dt\right)\left\|E_k^{(i)}\right\|_\infty + \lambda_{\text{tot}}\gamma\, dt\left\|E_k^{(j)}\right\|_\infty \\
&\quad + (F + CG)dt^2 + CM\left(\lambda_{\text{tot}}\gamma + 2\sqrt{\frac{3T\lambda_{\text{tot}}\gamma}{\varepsilon}}\right) dt^2 \\
&\quad + 2G\left(\lambda_{\text{tot}}\gamma + 2\sqrt{\frac{3T\lambda_{\text{tot}}\gamma}{\varepsilon}}\right) dt^2 + 2M\left(\lambda_{\text{tot}}\gamma + 2\sqrt{\frac{3T\lambda_{\text{tot}}\gamma}{\varepsilon}}\right)^2 dt^2 \\
&\quad + 2M\sqrt{\frac{3\kappa T\lambda_{\text{tot}}\gamma}{\varepsilon}}\, dt^2 + \widetilde{F}\, dt^2\,.
\end{aligned} \tag{3.137}$$

Grouping all of the prefactors of dt^2 into a single constant Z, we simplify (3.137) and write

$$\left\|E_{k+1}^{(i)}\right\|_\infty \leq \left(1 + (C + \lambda_{\text{tot}}\gamma)\, dt\right)\left\|E_k^{(i)}\right\|_\infty + \lambda_{\text{tot}}\gamma\, dt\left\|E_k^{(j)}\right\|_\infty + Z\, dt^2\,. \tag{3.138}$$

Let

$$E_k = \max_i\left\{\left\|E_k^{(i)}\right\|_\infty\right\} \tag{3.139}$$

denote the maximum error at time t_k. The maximum error at time 0 is $E_0 = 0$. Let $Y = (C + 2\lambda_{\text{tot}}\gamma)$. The maximum error at time t_{k+1} satisfies the bound

$$E_{k+1} \leq (1 + Y\, dt)E_k + Z\, dt^2\,. \tag{3.140}$$

With probability at least $1 - \frac{T}{dt}\left(\frac{2\varepsilon\, dt}{3T}\right) = 1 - \frac{2}{3\varepsilon} > 1 - \varepsilon$, the inequality (3.140) holds for all $k \in \{1, \ldots, \frac{T}{dt} - 1\}$. Therefore, with probability larger than $1 - \varepsilon$, the maximum error at time t_k satisfies

$$\begin{aligned} \|E_k\|_\infty &\leq Z\, dt^2 \sum_{k'=0}^{k-1} (1 + Y\, dt)^{k'} \\ &= Z\, dt^2 \left(\frac{(1 + Y\, dt)^k - 1}{(1 + Y dt) - 1} \right) \\ &\leq \frac{Z}{Y}\, dt \left(e^{Y\, dt}\right)^k \\ &\leq \frac{Z}{Y}\, dt \left(e^{Y dt}\right)^{\frac{T}{dt}} \\ &\leq \frac{Z e^{YT}}{Y}\, dt \end{aligned} \tag{3.141}$$

for all $k \in \{0, \ldots, \frac{T}{dt}\}$

Consider an arbitrary time $t' \in (t_k, t_{k+1})$. For some l, we have $t' \in [t_{k,l}, t_{k,l+1})$. It then follows that

$$\begin{aligned} E^{(i)}(t') &= \boldsymbol{N}^{(i)}(t') - \widetilde{\boldsymbol{N}}^{(i)}(t') \\ &= \boldsymbol{N}^{(i)}(t') - \boldsymbol{N}^{(i)}\left(t_{k,l}^+\right) + \boldsymbol{N}^{(i)}\left(t_{k,l}^+\right) - \boldsymbol{N}^{(i)}\left(t_{k,l}^-\right) \\ &\quad + \boldsymbol{N}^{(i)}\left(t_{k,l}^-\right) - \boldsymbol{N}^{(i)}(t_k) + \boldsymbol{N}^{(i)}(t_k) - \widetilde{\boldsymbol{N}}^{(i)}(t_k) \\ &\quad + \widetilde{\boldsymbol{N}}^{(i)}(t_k) - \widetilde{\boldsymbol{N}}^{(i)}(t')\,. \end{aligned} \tag{3.142}$$

Therefore,

$$\begin{aligned}
\left\|E^{(i)}(t')\right\|_\infty \leq & \left\|\boldsymbol{N}^{(i)}(t') - \boldsymbol{N}^{(i)}\left(t_{k,l}^{+}\right)\right\|_\infty + \left\|\boldsymbol{N}^{(i)}\left(t_{k,l}^{+}\right) - \boldsymbol{N}^{(i)}\left(t_{k,l}^{-}\right)\right\|_\infty \\
& + \left\|\boldsymbol{N}^{(i)}\left(t_{k,l}^{-}\right) - \boldsymbol{N}^{(i)}(t_k)\right\|_\infty + \left\|\boldsymbol{N}^{(i)}(t_k) - \widetilde{\boldsymbol{N}}^{(i)}(t_k)\right\|_\infty \\
& + \left\|\widetilde{\boldsymbol{N}}^{(i)}(t_k) - \widetilde{\boldsymbol{N}}^{(i)}(t')\right\|_\infty \\
\leq & \; G\left(t' - t_{k,l}\right) + M\gamma + \left\|A_{k,l}^{(i)}\right\|_\infty + \left\|E_k^{(i)}\right\|_\infty + \widetilde{G}(t' - t_k) \\
\leq & \; G\,dt + 4M\,dt^2 + (G\,dt + L_k M\gamma) + \frac{Ze^{YT}}{Y}\,dt + \widetilde{G}\,dt \\
\leq & \; \left(2G + 4M + \frac{Ze^{YT}}{Y} + \widetilde{G}\right)\,dt + M\left(\lambda_{\text{tot}}\gamma + 2\sqrt{\frac{3T\lambda_{\text{tot}}\gamma}{\varepsilon}}\right)\,dt\,, \\
\leq & \; \left(2G + 4M + \frac{Ze^{YT}}{Y} + \widetilde{G} + M\lambda_{\text{tot}}\gamma + 2M\sqrt{\frac{3T\lambda_{\text{tot}}\gamma}{\varepsilon}}\right)\,dt\,.
\end{aligned} \tag{3.143}$$

Because $dt < \gamma^{\frac{1}{4}}$, we have

$$\left\|E^{(i)}(t')\right\|_\infty < W\gamma^{\frac{1}{4}}\,, \tag{3.144}$$

where

$$W = \left(2G + 4M + \frac{Ze^{YT}}{Y} + \widetilde{G} + M\lambda_{\text{tot}}\gamma + 2M\sqrt{\frac{3T\lambda_{\text{tot}}\gamma}{\varepsilon}}\right)\,. \tag{3.145}$$

With the inequality

$$\gamma \leq \frac{\delta^4}{2W}\,, \tag{3.146}$$

it follows for all i and all times $t \in [0, T]$ that

$$\left\|E^{(i)}(t)\right\|_\infty \leq \delta \tag{3.147}$$

with probability larger than $1 - \varepsilon$.

□

3.7.3 Proof of High-Frequency Constant-Strength Approximation Theorem

In this appendix, we prove the HFCSA Theorem (see Theorem 3).

Theorem 3 (High-Frequency, Constant-Strength Approximation Theorem)**.** *Fix the relative interaction-frequency parameters l_{ij}, the interaction strength $\gamma > 0$, and a time T. Suppose that each local-dynamics function $g^{(i)}$ is Lipschitz continuous and bounded (see Section 3.2.3). Let $\varepsilon \in (0, 1]$, $\delta > 0$, and $\eta > 0$ be arbitrary but fixed constants. For sufficiently large λ_{tot}, each host microbiome abundance vector $\boldsymbol{N}^{(i)}(t)$ satisfies*

$$\left\| \boldsymbol{N}^{(i)} - \widetilde{\boldsymbol{N}} \right\|_{L^\infty[\eta,T]} < \delta \tag{3.35}$$

with probability larger than $1 - \varepsilon$, where

$$\begin{aligned} \frac{d\widetilde{\boldsymbol{N}}}{dt} &= \frac{1}{|H|} \sum_{j=1}^{|H|} g^{(j)} \left(\widetilde{\boldsymbol{N}} \right) , \\ \widetilde{\boldsymbol{N}}(0) &= \overline{\boldsymbol{N}}(0) . \end{aligned} \tag{3.36}$$

Proof. Each local-dynamics function $g^{(i)}$ is bounded (see Section 3.2.3), so there exists a constant M such that each entry of $\boldsymbol{N}^{(i)}(t)$ is nonnegative and

$$\left\| \boldsymbol{N}^{(i)}(t) \right\|_\infty \leq M \tag{3.148}$$

for each microbiome abundance vector $\boldsymbol{N}^{(i)}(t)$ and all times $t \geq 0$.

The approximate microbiome abundance vector $\widetilde{\boldsymbol{N}}t)$ is the mean of all $\boldsymbol{N}^{(i)}(t)$. The dynamics of $\widetilde{\boldsymbol{N}}(t)$ (3.36) is given by the mean of all hosts' local dynamics. Because each local-dynamics function $g^{(i)}$ is bounded, so is the mean of each $g^{(i)}$. Therefore, for all times $t \geq 0$, each entry of $\widetilde{\boldsymbol{N}}(t)$ is nonnegative and

$$\left\| \widetilde{\boldsymbol{N}}(t) \right\|_\infty \leq M . \tag{3.149}$$

We assume that each local-dynamics function $g^{(i)}$ is Lipschitz continuous. Therefore, there exists a constant C such that

$$\left\| g^{(i)}(\boldsymbol{x}) - g^{(i)}(\boldsymbol{y}) \right\|_\infty \leq C \left\| \boldsymbol{x} - \boldsymbol{y} \right\|_\infty \tag{3.150}$$

for each $g^{(i)}$ and all $\boldsymbol{x}, \boldsymbol{y} \in [0, M]^n$. Because each local-dynamics function $g^{(i)}$ is continuous (which follows from their Lipschitz continuity) and each microbiome abundance vector $\boldsymbol{N}^{(i)}(t)$ is in the compact region $[0, M]^n$, there exists a constant G such that

$$\left\| \frac{\mathrm{d}\boldsymbol{N}^{(i)}}{\mathrm{d}t} \right\|_\infty = \left\| g^{(i)}\left(\boldsymbol{N}^{(i)}\right) \right\|_\infty \leq G \tag{3.151}$$

for all $\boldsymbol{N}^{(i)}(t)$. For an interaction involving host $H^{(i)}$ that occurs at time t_I, we choose $\boldsymbol{N}^{(i)}(t_I) = \boldsymbol{N}^{(i)}\left(t_I^+\right)$. Therefore, $\boldsymbol{N}^{(i)}(t)$ is right-continuous at time t_I. It is usually not left-continuous at time t_I, so it is usually not left-differentiable at time t_I.[3] In such situations, the derivative that we use in (3.151) is a right derivative.

Define the Dirichlet energy

$$U(t) = \frac{1}{2} \sum_{i,j} \frac{\lambda_{ij}}{\lambda_{\mathrm{tot}}} \left\| \boldsymbol{N}^{(i)}(t) - \boldsymbol{N}^{(j)}(t) \right\|_2^2 , \tag{3.152}$$

which is nonnegative by construction. Let $\xi > 0$ be arbitrary but fixed. We will show that $U(t) \leq \xi$ with probability larger than $1 - \varepsilon$ for sufficiently large λ_{tot} and all times $t \in [\eta, T]$.

Between interactions,

$$\frac{\mathrm{d}U}{\mathrm{d}t} = \sum_{i.j} \frac{\lambda_{ij}}{\lambda_{\mathrm{tot}}} \left(g^{(i)}\left(\boldsymbol{N}^{(i)}\right) - g^{(j)}\left(\boldsymbol{N}^{(j)}\right) \right) \cdot \left(\boldsymbol{N}^{(i)} - \boldsymbol{N}^{(j)} \right) . \tag{3.153}$$

Therefore,

$$\begin{aligned} \left\| \frac{\mathrm{d}U}{\mathrm{d}t} \right\| &\leq \sum_{i.j} \frac{\lambda_{ij}}{\lambda_{\mathrm{tot}}} (2G) M \\ &= 4GM . \end{aligned} \tag{3.154}$$

Fix a $dt > 0$ such that T/dt is an integer and

$$dt < \max\left\{ \eta, \frac{\xi}{12GM}, \frac{\delta}{2CMe^{CT}} \right\} . \tag{3.155}$$

[3]The only situation where $\boldsymbol{N}^{(i)}(t)$ is left-continuous at time t_I occurs when the host $H^{(j)}$ with which $H^{(i)}$ interacts has a microbiome vector $\boldsymbol{N}^{(j)}(t_I^-) = \boldsymbol{N}^{(i)}(t_I^-)$.

Let $t_k = k\,dt$. There are only finitely many t_k in the interval $[0, T]$. The probability that an interaction occurs precisely at any of these t_k is 0. Therefore, for the remainder of this proof, we only consider interactions that occur at times $t \neq t_k$ for any k. Under this assumption,

$$U(t_k^-) = U(t_k) \tag{3.156}$$

for each $t_k \in [0, T]$.

We now consider how the Dirichlet energy $U(t)$ changes over an interval $[t_k, t_{k+1}]$. Let L_k be the number of interactions that occur in (t_k, t_{k+1}). (This interval is open because no interactions occur at any of the t_k.) We denote the associated ordered set of interactions by $\{t_{k,l}\}_{l=1}^{L_k}$. Additionally, we let $t_{k,0} = t_k$ and $t_{k,L_k+1} = t_{k+1}$, and we define $dt_{k,l} = t_{k,l} - t_{k,l-1}$. For $l \in \{1, \ldots, L_k\}$, let

$$W_{k,l} = U\left(t_{k,l}^+\right) - U\left(t_{k,l}^-\right) . \tag{3.157}$$

The difference $W_{k,l}$ is the change in U due to an interaction at time $t_{k,l}$.

We decompose the change in $U(t)$ over the interval $[t_k, t_k + 1]$ by writing

$$\begin{aligned} U(t_{k+1}) - U(t_k) &= \sum_{l=1}^{L_k+1} \left[U\left(t_{k,l}^-\right) - U\left(t_{k,l-1}^-\right)\right] \\ &= \sum_{l=1}^{L_k+1} \left[U\left(t_{k,l}^-\right) - U\left(t_{k,l-1}^+\right)\right] + \sum_{l=2}^{L_k+1} \left[U\left(t_{k,l-1}^+\right) - U\left(t_{k,l-1}^-\right)\right] \\ &= \sum_{l=1}^{L_k+1} \left[U\left(t_{k,l}^-\right) - U\left(t_{k,l-1}^+\right)\right] + \sum_{l=1}^{L_k} W_{k,l} \, . \end{aligned} \tag{3.158}$$

The magnitude of the first sum in (3.158) has the upper bound

$$\begin{aligned} \left\| \sum_{l=1}^{L_k+1} \left[U\left(t_{k,l}^-\right) - U\left(t_{k,l-1}^+\right)\right] \right\| &\leq \sum_{l=1}^{L_k+1} \left\| \left[U\left(t_{k,l}^-\right) - U\left(t_{k,l-1}^+\right)\right] \right\| \\ &\leq \sum_{l=1}^{L_k+1} 4GM\, dt_{k,l} \\ &= 4GM\, dt \, . \end{aligned} \tag{3.159}$$

We now consider the sum $\sum_{l=1}^{L_k} W_{k,l}$. If the interaction at time $t_{k,l}$ is between hosts $H^{(i)}$ and $H^{(j)}$, then

$$\begin{aligned} W_{k,l} &= \frac{\lambda_{ij}}{\lambda_{\text{tot}}} \left[\left\| \boldsymbol{N}^{(i)}\left(t_I^+\right) - \boldsymbol{N}^{(j)}\left(t_I^+\right) \right\|_2^2 - \left\| \boldsymbol{N}^{(i)}\left(t_I^-\right) - \boldsymbol{N}^{(j)}\left(t_I^-\right) \right\|_2^2 \right] \\ &= \frac{\lambda_{ij}}{\lambda_{\text{tot}}} \left[\begin{array}{l} \left\| (1-\gamma)\boldsymbol{N}^{(i)}\left(t_I^-\right) + \gamma \boldsymbol{N}^{(j)}\left(t_I^-\right) - (1-\gamma)\boldsymbol{N}^{(j)}\left(t_I^-\right) - \gamma \boldsymbol{N}^{(i)}\left(t_I^-\right) \right\|_2^2 \\ \quad - \left\| \boldsymbol{N}^{(i)}\left(t_I^-\right) - \boldsymbol{N}^{(j)}\left(t_I^-\right) \right\|_2^2 \end{array} \right] \\ &= \frac{\lambda_{ij}}{\lambda_{\text{tot}}} \left[(1-2\gamma)^2 \left\| \boldsymbol{N}^{(i)}\left(t_I^-\right) - \boldsymbol{N}^{(j)}\left(t_I^-\right) \right\|_2^2 - \left\| \boldsymbol{N}^{(i)}\left(t_I^-\right) - \boldsymbol{N}^{(j)}\left(t_I^-\right) \right\|_2^2 \right] \\ &= -4\gamma(1-\gamma) \frac{\lambda_{ij}}{\lambda_{\text{tot}}} \left\| \boldsymbol{N}^{(i)}\left(t_I^-\right) - \boldsymbol{N}^{(j)}\left(t_I^-\right) \right\|_2^2 . \end{aligned} \tag{3.160}$$

Regardless of which hosts interact, each $W_{k,l}$ is always nonpositive. We now show by contradiction that the Dirichlet energy $U(t) \leq \xi/3$ with probability at least $1 - \frac{\varepsilon\, dt}{2T}$ for sufficiently large total-interaction-frequency parameter λ_{tot} and some time $t \in [t_k, t_{k+1}]$.

Suppose that $U(t) > \xi/3$ for all times $t \in [t_k, t_{k+1}]$. For all $t \in [t_k, t_{k+1}]$, there are then some i and j such that

$$\frac{\lambda_{ij}}{\lambda_{\text{tot}}} \left\| \boldsymbol{N}^{(i)}(t) - \boldsymbol{N}^{(j)}(t) \right\|_2^2 > \frac{\xi}{3|H|^2} . \tag{3.161}$$

Therefore, at each time $t_{k,l}^-$ immediately before an interaction, there is some i and j such that

$$\frac{\lambda_{ij}}{\lambda_{\text{tot}}} \left\| \boldsymbol{N}^{(i)}\left(t_{k,l}^-\right) - \boldsymbol{N}^{(j)}\left(t_{k,l}^-\right) \right\|_2^2 > \frac{\xi}{3|H|^2} . \tag{3.162}$$

Let

$$l_{\min} = \min_{i,j} \left\{ l_{ij} \mid l_{ij} > 0 \right\} . \tag{3.163}$$

An interaction at time $t_{k,l}$ occurs between hosts $H^{(i)}$ and $H^{(j)}$ with probability $l_{ij} \geq l_{\min}$. Therefore, regardless of the previous interactions, each $W_{k,l}$ satisfies

$$W_{k,l} < -\frac{4\gamma(1-\gamma)\xi}{3|H|^2} \tag{3.164}$$

with probability at least $l_{\min}$.

We define a random variable $\mathcal{W}$ such that

$$\Pr\left(\mathcal{W}=0\right)=1-l_{\min}\,, \tag{3.165}$$
$$\Pr\left(\mathcal{W}=-\frac{4\gamma(1-\gamma)\xi}{3|H|^2}\right)=l_{\min}\,.$$

For all $w\leq 0$, we have

$$\Pr\left(\sum_{l=1}^{L_k}W_{k,l}<w\right)\geq\Pr\left(\sum_{l=1}^{L_k}\mathcal{W}<w\right)\,. \tag{3.166}$$

We seek to bound

$$\sum_{l=1}^{L_k}\mathcal{W}\,. \tag{3.167}$$

The random variable $\mathcal{W}$ is stochastic, so we can only find a bound for (3.167) that holds with some probability. The first two moments of $\mathcal{W}$ are

$$\mathbb{E}\left[\mathcal{W}\right]=-l_{\min}\frac{4\gamma(1-\gamma)\xi}{3|H|^2}\,, \tag{3.168}$$
$$\mathbb{E}\left[\mathcal{W}^2\right]=l_{\min}\left(\frac{4\gamma(1-\gamma)\xi}{3|H|^2}\right)^2\,.$$

The expectation of the sum (3.167) is

$$\begin{aligned}\mathbb{E}\left[\sum_{l=1}^{L_k}\mathcal{W}\right]&=\mathbb{E}\left[\,\mathbb{E}\left[\sum_{l=1}^{L_k}\mathcal{W}\,\middle|\,L_k\right]\right] \\ &=\mathbb{E}\left[L_k\left(-l_{\min}\frac{4\gamma(1-\gamma)\xi}{3|H|^2}\right)\right] \\ &=-\lambda_{\text{tot}}dt\,l_{\min}\frac{4\gamma(1-\gamma)\xi}{3|H|^2}\,.\end{aligned} \tag{3.169}$$

Applying the law of total variance, the variance of the sum (3.167) is

$$\begin{aligned}\text{Var}\left[\sum_{l=1}^{L_k}\mathcal{W}\right]&=\mathbb{E}\left[\,\text{Var}\left[\sum_{l=1}^{L_k}\mathcal{W}\,\middle|\,L_k\right]\right]+\text{Var}\left[\,\mathbb{E}\left[\sum_{l=1}^{L_k}\mathcal{W}\,\middle|\,L_k\right]\right] \\ &=\mathbb{E}\left[L_k\text{Var}\left[\mathcal{W}\right]\right]+\text{Var}\left[\,L_k\mathbb{E}\left[\mathcal{W}\right]\right] \\ &=\lambda_{\text{tot}}dt\,\text{Var}\left[\mathcal{W}\right]+\lambda_{\text{tot}}dt\left(\mathbb{E}\left[\mathcal{W}\right]\right)^2 \\ &=\lambda_{\text{tot}}dt\,\mathbb{E}\left[\mathcal{W}^2\right] \\ &=\lambda_{\text{tot}}dt\,l_{\min}\left(\frac{4\gamma(1-\gamma)\xi}{3|H|^2}\right)^2\,.\end{aligned} \tag{3.170}$$

Let $\alpha = \sqrt{\frac{2T}{\varepsilon\, dt}}$. By Chebyshev's inequality,

$$\Pr\left(\sum_{l=1}^{L_k} \mathcal{W} \geq \mathbb{E}\left[\sum_{l=1}^{L_k} \mathcal{W}\right] + \alpha \sqrt{\operatorname{Var}\left[\sum_{l=1}^{L_k} \mathcal{W}\right]} \right) \leq \frac{1}{\alpha^2}\,, \tag{3.171}$$

$$\Pr\left(\sum_{l=1}^{L_k} \mathcal{W} \geq -\lambda_{\text{tot}} dt\, l_{\min} \frac{4\gamma(1-\gamma)\xi}{3|H|^2} + \sqrt{\frac{2T\lambda_{\text{tot}} l_{\min}}{\varepsilon}} \frac{4\gamma(1-\gamma)\xi}{3|H|^2} \right) \leq \frac{\varepsilon\, dt}{2T}\,,$$

$$\Pr\left(\sum_{l=1}^{L_k} \mathcal{W} < -\left(\lambda_{\text{tot}} dt\, l_{\min} - \sqrt{\frac{2T\lambda_{\text{tot}} l_{\min}}{\varepsilon}} \right) \frac{4\gamma(1-\gamma)\xi}{3|H|^2} \right) \geq 1 - \frac{\varepsilon\, dt}{2T}\,.$$

Using the bound (3.166) in (3.171) gives

$$\Pr\left(\sum_{l=1}^{L_k} W_{k,l} < -\left(\lambda_{\text{tot}} dt\, l_{\min} - \sqrt{\frac{2T\lambda_{\text{tot}} l_{\min}}{\varepsilon}} \right) \frac{4\gamma(1-\gamma)\xi}{3|H|^2} \right) > 1 - \frac{\varepsilon\, dt}{2T}\,, \tag{3.172}$$

$$\Pr\left(\sum_{l=1}^{L_k} W_{k,l} < -\sqrt{\lambda_{\text{tot}}} \left(\sqrt{\lambda_{\text{tot}}} dt\, l_{\min} - \sqrt{\frac{2T l_{\min}}{\varepsilon}} \right) \frac{4\gamma(1-\gamma)\xi}{3|H|^2} \right) > 1 - \frac{\varepsilon\, dt}{2T}\,.$$

By choosing sufficiently large λ_{tot}, we can make $\sqrt{\lambda_{\text{tot}}} \left(\sqrt{\lambda_{\text{tot}}} dt\, l_{\min} - \sqrt{\frac{2T l_{\min}}{\varepsilon}} \right) \frac{4\gamma(1-\gamma)\xi}{3|H|^2}$ arbitrarily large. In particular, we choose a sufficiently large λ_{tot} so that

$$\sqrt{\lambda_{\text{tot}}} \left(\sqrt{\lambda_{\text{tot}}} dt l_{\min} - \sqrt{\frac{2T l_{\min}}{\varepsilon}} \right) \frac{4\gamma(1-\gamma)\xi}{3|H|^2} \geq U(0) + 4GMT + 4GM\, dt\,. \tag{3.173}$$

Therefore,

$$\Pr\left(\sum_{l=1}^{L_k} W_{k,l} < -\left[U(0) + 4GMT + 4GM\, dt \right] \right) > 1 - \frac{\varepsilon\, dt}{2T}\,. \tag{3.174}$$

Local dynamics can cause the Dirichlet energy $U(t)$ to change at a rate of at most $4GM$ per unit time (see (3.154)). Additionally, because each $W_{k,l}$ is nonpositive, interactions cannot cause U to increase. Therefore, for any time $t \in [0, T]$, we have $U(t) < U(0) + 4GMT$. Combining this upper bound, the bound (3.174), and the decomposition of $U(t_{k+1}) - U(t_k)$ in (3.158) yields

$$U(t_{k+1}) \leq U(t_k) + \left\| \sum_{l=1}^{L_k+1} \left[U\left(t_{k,l}^{-}\right) - U\left(t_{k,l-1}^{+}\right) \right] \right\| + \left\| \sum_{l=1}^{L_k} W_{k,l} \right\| \tag{3.175}$$

$$< U(0) + 4GMT + 4GM\, dt - [U(0) + 4GMT + 4GM\, dt]$$

$$< 0 \tag{3.176}$$

with probability larger than $1 - \frac{\varepsilon\, dt}{2T}$. However, $U(t)$ is always nonnegative. Therefore, with probability larger than $1 - \frac{\varepsilon\, dt}{2T}$, we have a contradiction and there is some time $t' \in [t_k, t_{k+1}]$ such that $U(t') \le \xi/3$. For some l', we have $t' \in [t_{k,l'}, t_{k,l'+1})$. Therefore,

$$U(t_{k+1}) = U(t') + U(t_{k,l'+1}) - U(t') + \sum_{l=l'+2}^{L_k+1} \left[U\left(t_{k,l}^-\right) - U\left(t_{k,l-1}^+\right)\right] + \sum_{l=l'+1}^{L_k} W_{k,l} \tag{3.177}$$

$$\begin{aligned} &\le U(t') + \|U(t_{k,l'+1}) - U(t')\| + \sum_{l=l'+2}^{L_k+1} \left\|U\left(t_{k,l}^-\right) - U\left(t_{k,l-1}^+\right)\right\| + \left\|\sum_{l=l'+1}^{L_k} W_{k,l}\right\| \\ &\le \frac{\xi}{3} + 4GM\, dt_{k,l'+1} + \sum_{l=l'+2}^{L_k+1} 4GM\, dt_{k,l} \\ &\le \frac{\xi}{3} + 4GM\, dt\,. \end{aligned} \tag{3.178}$$

Recall that $dt < \frac{\xi}{12GM}$ (see (3.155)). We have

$$U(t_{k+1}) < \frac{\xi}{3} + \frac{\xi}{3} = \frac{2\xi}{3}\,.$$

With probability larger than $1 - \frac{\varepsilon\, dt}{2T}\left(\frac{T}{dt}\right) = 1 - \frac{\varepsilon}{2} > 1 - \varepsilon$, every $U(t_k)$ except $U(0)$ satisfies $U(t_k) < 2\xi/3$. Consider an arbitrary time $t' \in [t_k, t_{k+1}]$, where $k \ge 1$. For some l', we have $t' \in [t_{k,l'}, t_{k,l'+1})$. Therefore,

$$U(t') = U(t') - U(t_{k,l'}) + \sum_{l=1}^{l'} \left[U\left(t_{k,l}^-\right) - U\left(t_{k,l-1}^+\right)\right] + \sum_{l=1}^{l'} W_{k,l} + U(t_k)\,, \tag{3.179}$$

which implies that

$$U(t') \le \|U(t') - U(t_{k,l'})\| + \sum_{l=1}^{l'} \left\|U\left(t_{k,l}^-\right) - U\left(t_{k,l-1}^+\right)\right\| + U(t_k) \tag{3.180}$$

$$\begin{aligned} &\le 4GM\, dt_{k,l'+1} + \sum_{l=1}^{l'} 4GM\, dt_{k,l} + U(t_k) \\ &\le 4GM\, dt + U(t_k) \\ &< \frac{\xi}{3} + \frac{2\xi}{3} \\ &= \xi\,. \end{aligned}$$

With probability larger than $1-\varepsilon$, we have $U(t) < \xi$ for all times $t \in [dt, T]$. Because ξ is arbitrary, by choosing a sufficiently large total-interaction-frequency parameter λ_{tot}, we can make $U(t)$ arbitrarily small over the interval $[dt, T]$ with probability larger than $1-\varepsilon$. Assume that $U(t) < \xi$ for all $t \in [dt, T]$. We can then bound the maximum difference between any microbiome abundance vectors $\boldsymbol{N}^{(i)}(t)$ and $\boldsymbol{N}^{(j)}(t)$. In particular, if $H^{(i)}$ and $H^{(j)}$ are adjacent, then

$$\begin{aligned} \frac{\lambda_{ij}}{\lambda_{\text{tot}}} \left\| \boldsymbol{N}^{(i)}(t) - \boldsymbol{N}^{(j)}(t) \right\|_2^2 &< \xi \,, \\ \left\| \boldsymbol{N}^{(i)}(t) - \boldsymbol{N}^{(j)}(t) \right\|_2 &< \sqrt{l_{\max} \xi} \,, \end{aligned} \tag{3.181}$$

where $l_{\max} = \max_{i,j} l_{ij}$. A shortest path between any two hosts in the interaction network has length at most $|H| - 1$. Therefore, the 2-norm of the difference between each pair of microbiome abundance vectors $\boldsymbol{N}^{(i)}(t)$ and $\boldsymbol{N}^{(j)}(t)$ satisfies

$$\left\| \boldsymbol{N}^{(i)}(t) - \boldsymbol{N}^{(j)}(t) \right\|_2 < (|H| - 1) \sqrt{l_{\max} \xi} \,. \tag{3.182}$$

The inequality (3.182) guarantees that the difference between each microbiome abundance vector $\boldsymbol{N}^{(i)}(t)$ and the mean microbiome abundance vector $\overline{\boldsymbol{N}}(t)$ satisfies

$$\begin{aligned} \left\| \boldsymbol{N}^{(i)}(t) - \overline{\boldsymbol{N}}(t) \right\|_2 &< (|H| - 1) \sqrt{l_{\max} \xi} \,, \\ \left\| \boldsymbol{N}^{(i)}(t) - \overline{\boldsymbol{N}}(t) \right\|_\infty &< (|H| - 1) \sqrt{l_{\max} \xi} \,. \end{aligned} \tag{3.183}$$

We now bound the magnitude of the difference between the mean microbiome abundance vector $\overline{\boldsymbol{N}}(t)$ and the approximate microbiome abundance vector $\widetilde{\boldsymbol{N}}(t)$ for all times $t \in [0, T]$. We refer to $\left\| \overline{\boldsymbol{N}}(t) - \widetilde{\boldsymbol{N}}(t) \right\|_\infty$ as the *approximation–mean error* at time t. At time 0, the approximation–mean error $\left\| \overline{\boldsymbol{N}}(0) - \widetilde{\boldsymbol{N}}(0) \right\|_\infty = 0$ by construction. We bound the

approximation–mean error by first bounding its time derivative, which satisfies

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t}\left\|\overline{\boldsymbol{N}}-\widetilde{\boldsymbol{N}}\right\|_\infty &\leq \left\|\frac{\mathrm{d}}{\mathrm{d}t}\left(\overline{\boldsymbol{N}}-\widetilde{\boldsymbol{N}}\right)\right\|_\infty \\
&= \left\|\frac{1}{|H|}\sum_{j=1}^{|H|} g^{(j)}\left(\boldsymbol{N}^{(j)}\right)-\frac{1}{|H|}\sum_{j=1}^{|H|} g^{(j)}\left(\widetilde{\boldsymbol{N}}\right)\right\|_\infty \\
&\leq \frac{1}{|H|}\sum_{j=1}^{|H|}\left\|g^{(j)}\left(\boldsymbol{N}^{(j)}\right)-g^{(j)}\left(\widetilde{\boldsymbol{N}}\right)\right\|_\infty \\
&\leq \frac{1}{|H|}\sum_{j=1}^{|H|} C\left\|\boldsymbol{N}^{(j)}-\widetilde{\boldsymbol{N}}\right\|_\infty \\
&\leq \frac{1}{|H|}\sum_{j=1}^{|H|} C\left\|\boldsymbol{N}^{(j)}-\widetilde{\boldsymbol{N}}\right\|_\infty \\
&\leq \frac{1}{|H|}\sum_{j=1}^{|H|} C\left(\left\|\boldsymbol{N}^{(j)}-\overline{\boldsymbol{N}}\right\|_\infty+\left\|\overline{\boldsymbol{N}}-\widetilde{\boldsymbol{N}}\right\|_\infty\right) \\
&\leq C\left\|\overline{\boldsymbol{N}}-\widetilde{\boldsymbol{N}}\right\|_\infty+\frac{1}{|H|}\sum_{j=1}^{|H|} C\left\|\boldsymbol{N}^{(j)}-\overline{\boldsymbol{N}}\right\|_\infty .
\end{aligned} \tag{3.184}$$

For times $t \in [dt, T]$, each $\left\|\boldsymbol{N}^{(j)}(t)-\overline{\boldsymbol{N}}(t)\right\|_\infty < (|H|-1)\sqrt{l_{\max}\xi}$. Therefore, on this interval,

$$\frac{\mathrm{d}}{\mathrm{d}t}\left\|\overline{\boldsymbol{N}}-\widetilde{\boldsymbol{N}}\right\|_\infty < C\left\|\overline{\boldsymbol{N}}-\widetilde{\boldsymbol{N}}\right\|_\infty+(|H|-1)\sqrt{l_{\max}\xi}\,. \tag{3.185}$$

For $t \in [0, dt]$, we obtain a weaker bound for the time derivative of the approximation–mean error:

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t}\left\|\overline{\boldsymbol{N}}-\widetilde{\boldsymbol{N}}\right\|_\infty &\leq \frac{1}{|H|}\sum_{j=1}^{|H|} C\left\|\boldsymbol{N}^{(j)}-\widetilde{\boldsymbol{N}}\right\|_\infty \\
&\leq \frac{1}{|H|}\sum_{j=1}^{|H|} CM \\
&= CM\,.
\end{aligned} \tag{3.186}$$

Therefore, for times $t \in [0, dt]$, the approximation–mean error satisfies

$$\begin{aligned} \left\| \overline{\boldsymbol{N}}(t) - \widetilde{\boldsymbol{N}}(t) \right\|_\infty &\leq \left\| \overline{\boldsymbol{N}}(0) - \widetilde{\boldsymbol{N}}(0) \right\|_\infty + CMt \\ &\leq CM\, dt \\ &\leq \frac{\delta}{2e^{CT}}\,. \end{aligned} \tag{3.187}$$

For times $t \in [dt, T]$, we construct a function $u(t)$ that gives an upper bound of the approximation–mean error. This function $u(t)$ is the solution of the dynamical system

$$\begin{aligned} \frac{\mathrm{d}u}{\mathrm{d}t} &= Cu + (|H| - 1)\sqrt{l_{\max}\xi}\,, \\ u(dt) &= \frac{\delta}{2e^{CT}}\,. \end{aligned} \tag{3.188}$$

Using an integrating factor to solve for $u(t)$ yields

$$\begin{aligned} e^{-Ct}\frac{\mathrm{d}u}{\mathrm{d}t} &= e^{-Ct}Cu + e^{-Ct}(|H| - 1)\sqrt{l_{\max}\xi}\,, \\ \frac{\mathrm{d}}{\mathrm{d}t}\left(e^{-Ct}u\right) &= e^{-Ct}(|H| - 1)\sqrt{l_{\max}\xi}\,, \\ e^{-Ct}u(t) &= e^{-Cdt}\left(\frac{\delta}{2e^{CT}}\right) + \int_{dt}^{t} e^{-Ct'}(|H| - 1)\sqrt{l_{\max}\xi}\;\mathrm{d}t'\,, \\ u(t) &= e^{C(t-dt)}\left(\frac{\delta}{2e^{CT}}\right) + \frac{(|H| - 1)\sqrt{l_{\max}\xi}}{C}\left(e^{C(t-dt)} - 1\right)\,. \end{aligned} \tag{3.189}$$

For $t \in [dt, T]$, the function $u(t)$ satisfies

$$\begin{aligned} u(t) &\leq e^{CT}\left(\frac{\delta}{2e^{CT}}\right) + \frac{(|H| - 1)\sqrt{l_{\max}\xi}}{C}e^{CT} \\ &= \frac{\delta}{2} + \frac{(|H| - 1)\sqrt{l_{\max}\xi}}{C}e^{CT}\,. \end{aligned} \tag{3.190}$$

Therefore,

$$\left\| \overline{\boldsymbol{N}}(t) - \widetilde{\boldsymbol{N}}(t) \right\|_\infty \leq \frac{\delta}{2} + \frac{(|H| - 1)\sqrt{l_{\max}\xi}}{C}e^{CT}\,. \tag{3.191}$$

As we discussed previously (see (3.180)), we can make ξ arbitrarily small by choosing a sufficiently large λ_{tot}. Therefore, we make ξ sufficiently small so that

$$(|H| - 1)\sqrt{l_{\max}\xi} < \max\left\{\frac{\delta}{4}, \frac{\delta C}{4e^{CT}}\right\}\,. \tag{3.192}$$

Using the bound (3.192) in (3.191) and (3.183) yields

$$\begin{aligned} \left\| \overline{\boldsymbol{N}}(t) - \widetilde{\boldsymbol{N}}(t) \right\|_\infty &< \frac{\delta}{2} + \frac{\delta C}{4e^{CT}} \left(\frac{e^{CT}}{C} \right) = \frac{3\delta}{4}, \\ \left\| \boldsymbol{N}^{(i)}(t) - \overline{\boldsymbol{N}}(t) \right\|_\infty &< \frac{\delta}{4} \end{aligned} \tag{3.193}$$

for every $t \in [dt, T]$ with probability larger than $1 - \varepsilon$. Because $dt \leq \eta$, for all times $t \in [\eta, T]$, the difference between each microbiome abundance vector $\boldsymbol{N}^{(i)}(t)$ and the approximate microbiome abundance vector $\widetilde{\boldsymbol{N}}(t)$ satisfies

$$\begin{aligned} \left\| \boldsymbol{N}^{(i)}(t) - \widetilde{\boldsymbol{N}}(t) \right\|_\infty &\leq \left\| \boldsymbol{N}^{(i)}(t) - \overline{\boldsymbol{N}}(t) \right\|_\infty + \left\| \overline{\boldsymbol{N}}(t) - \widetilde{\boldsymbol{N}}(t) \right\|_\infty \\ &< \frac{\delta}{4} + \frac{3\delta}{4} = \delta \end{aligned} \tag{3.194}$$

with probability larger than $1 - \varepsilon$.

□

CHAPTER 4

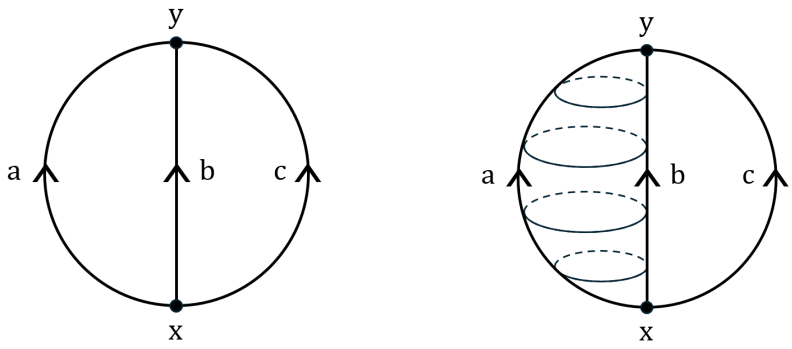
Background on Persistent Homology

In this chapter, we describe the mathematical background on *persistent homology* (PH) [OPT17] that we need in Chapter 5. In Chapter 5, we use *topological data analysis* (TDA) [EH10, DW22] to study "holes" in resource coverage. To do this, we apply PH, which is one of the main tools in TDA. PH uses ideas from algebraic topology to (1) identify clusters and holes in a data set and (2) measure their persistences across different scales.

In Section 4.1, we describe the theory of *homology*. In Section 4.2, we outline how to compute PH for a point cloud. Section 4.2 is adapted from the background section of [HJJ24], which was led jointly by Abigail Hickok, Benjamin Jarman, Jiajie Luo, and me and was co-authored with Mason A. Porter.

4.1 Homology

Before introducing persistent homology (PH), it is helpful to build some intuition for homology theory [Hat02], which provides the mathematical foundation for PH. Homology theory is a branch of algebraic topology that characterizes a topological space by its "holes". Let X be a topological space. For $k \geq 1$, the kth *homology group* $H_k(X, \mathbb{Z})$ of X with integer coefficients is the Abelian group that represents the "k-dimensional holes" of the space X. The rank of this group gives the number of such holes. For $k = 0$, the rank of $H_0(X, \mathbb{Z})$ gives the number of connected components of X. If there are multiple connected components, the space between these components is considered a 0-dimensional (0D) hole. The rank of



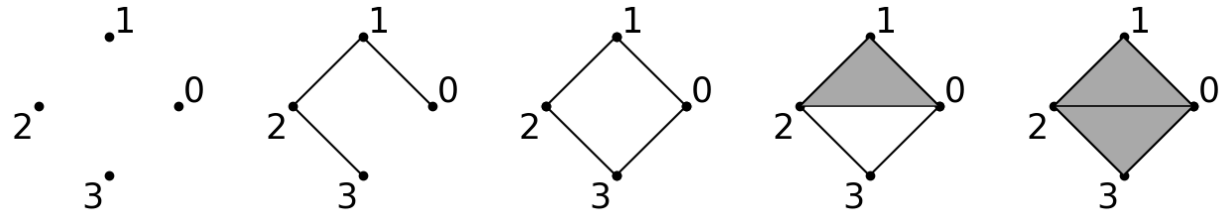
(a) X_1 (b) X_2

Figure 4.1: An example of two topological spaces X_1 and X_2. [These figures are adapted from figures in Chapter 2 of [Hat02].]

$H_1(X, \mathbb{Z})$ gives the number of "loops" in X that cannot be contracted to a point. The rank of $H_2(X, \mathbb{Z})$ gives the number of regions that are bounded by a 2-dimensional surface, and so on for higher values of k.

In Figure 4.1, we show two examples of topological spaces, X_1 and X_2. We adapt these figures from Chapter 2 of [Hat02]. The space X_1 is a graph with three edges a, b, and c that connect two nodes x and y. We assign directions to the edges from x to y. The space X_2 has the same two nodes and the same three edges. Additionally, X_2 has two surfaces, which each have the boundary $a - b$, where $a - b$ is the loop that goes from x to y along edge a and back to x along edge b.

Both of the spaces X_1 and X_2 have a single connected component. Therefore, $H_0(X_1, \mathbb{Z})$ and $H_0(X_2, \mathbb{Z})$ each have rank 1. In X_1, the two loops $a - b$ and $b - c$ are a basis for all other loops. One can construct any other loop from a linear combination of these two loops. For example, $a - c = (a - b) + (b - c)$. Therefore, $H_1(X_1, \mathbb{Z})$ has rank 2, indicating that there are two 1-dimensional holes. For X_2, however, the loop $a - b$ is filled in, so $H_1(X_2, \mathbb{Z})$ has rank 1. The space X_1 has no 2-dimensional holes, so $H_2(X_1, \mathbb{Z})$ has rank 0. The space X_2 has one 2-dimensional hole. The surface of this 2-dimensional hole is the union of the



(a) $\mathcal{K}_0$ (b) $\mathcal{K}_1$ (c) $\mathcal{K}_2$ (d) $\mathcal{K}_3$ (e) $\mathcal{K}_4$

Figure 4.2: An example of a filtration. The simplicial complex $\mathcal{K}_i$ has the associated filtration-parameter value i. [This figure appeared originally in [HNP22] and is used with permission.]

two surfaces with boundary $a - b$. Therefore, $H_2(X_2, \mathbb{Z})$ has rank 0. For $k \geq 3$, the groups $H_k(X_1, \mathbb{Z})$ and $H_k(X_2, \mathbb{Z})$ both have rank 0.

4.2 Persistent Homology for Point Clouds

We review relevant mathematical background on our application of PH in Chapter 5. See [OPT17, EH10, DW22] for more thorough discussions. To compute PH, we begin by constructing a *filtered simplicial complex* (which we will call a *filtration*) from a *point cloud*, which is a finite collection $X = \{x_i\}_{i=1}^n$ of points in a metric space (M, d). A *simplicial complex* is a combinatorial description of a topological space. It is a collection of vertices, edges, triangles, and higher-dimensional simplices with certain requirements on simplex boundaries and pairwise simplex intersections. A filtration is a nested sequence $\mathcal{K}_{\alpha_0} \subseteq \mathcal{K}_{\alpha_1} \subseteq \cdots \subseteq \mathcal{K}_{\alpha_n}$ of simplicial complexes, where $\alpha_0 < \alpha_1 < \cdots < \alpha_n$. We show an example of a filtration in Figure 4.2.

Two of the most common constructions are the Čech filtration and the Vietoris–Rips (VR) filtration [OPT17]. For $r > 0$, the *Čech complex* $\check{\mathrm{C}}_r(X, M, d)$ *at filtration parameter* r is the simplicial complex that has a simplex with vertices $[x_{i_0}, \ldots, x_{i_k}]$ if the intersection $\bigcap_j B(x_{i_j}, r)$ is nonempty, where $B(x, r) := \{y \in M \mid d(x, y) \leq r\}$. That is, $\check{\mathrm{C}}_r(X, M, d)$ is

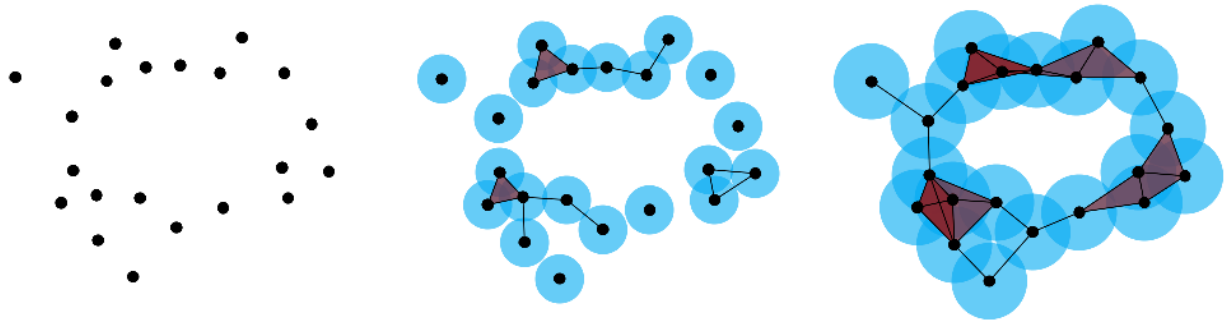


Figure 4.3: Illustration of a Čech filtration for a point cloud X that we sample from an annulus. [This figure appeared originally in [HJJ24]. We generated this figure using [AS11].]

the nerve of the closed balls $\{B(x_i, r)\}_{x_i \in X}$. By the Nerve Theorem [Bor48], $\check{\mathrm{C}}_r(X, M, d)$ is topologically equivalent (more precisely, it is homotopy-equivalent) to the union $\bigcup_i B(x_i, r)$ of balls (i.e., the *r-coverage* of X) in M whenever the balls $B(x_i, r)$ are convex.[1] This implies that $\bigcup_i B(x_i, r)$ and $\check{\mathrm{C}}_r(X, M, d)$ have the same homology (i.e., the same set of holes). A *Čech filtration* is a nested sequence of Čech complexes for increasing filtration parameter r. In Figure 4.3, we show an example of a Čech filtration.

In practice, it is uncommon to use Čech filtrations because they are difficult to compute. A *Vietoris–Rips (VR) complex* $\mathrm{VR}_r(X, M, d)$ is an approximation of a Čech complex that is faster to compute because it is only necessary to calculate pairwise distances between points. The VR complex at filtration parameter r has a simplex with vertices $[x_{i_0}, \ldots, x_{i_k}]$ if $d(x_{i_j}, x_{i_\ell}) < 2r$ for all j and ℓ. A *VR filtration* is a nested sequence of VR complexes for increasing filtration parameter r. A VR filtration approximates a Čech filtration in the sense that

$$\check{\mathrm{C}}_r(X, M, d) \subseteq \mathrm{VR}_r(X, M, d) \subseteq \check{\mathrm{C}}_{\sqrt{2}r}(X, M, d) \tag{4.1}$$

for all r. The complexes $\mathrm{VR}_r(X, M, d)$ and $\check{\mathrm{C}}_r(X, M, d)$ have the same set of edges for all r.

In Chapter 5, we will use a *distance* that captures the time cost of accessing a resource.

[1] This condition is satisfied for all r when (M, d) is Euclidean, but it is not always satisfied for non-Euclidean metric spaces.

The filtration level t represents this time cost. We use this in our notation for weighted versions of the Čech and VR filtration [ACG19]. Given a point cloud $X = \{x_1, \ldots, x_n\}$ in a metric space (M, d) and associated weights $\{w_1, \ldots, w_n\}$, the *radius function* at x_i is

$$r_{x_i}(t) := \begin{cases} -\infty\,, & t < w_i \\ t - w_i\,, & \text{otherwise}\,. \end{cases} \tag{4.2}$$

The closed ball $B(x_i, r_{x_i}(t))$ has no points for $t < w_i$; for $t \geq w_i$, the radius grows linearly with t, which is the filtration parameter. The *weighted Čech complex* $\check{\mathrm{C}}_t^{\mathrm{weighted}}(X, M, d, \{w_i\})$ at filtration parameter t is the simplicial complex that has a simplex with vertices $[x_{i_0}, \ldots, x_{i_k}]$ if the intersection $\bigcap_j B(x_{i_j}, r_{x_{i_j}}(t))$ is nonempty. That is, $\check{\mathrm{C}}_t^{\mathrm{weighted}}(X, M, d, \{w_i\})$ is the nerve of $\{B(x_i, r_{x_i}(t))\}_{x_i \in X}$. Like the unweighted Čech complex, the weighted Čech complex is homotopy-equivalent to the union $\bigcup_i B(x_i, r_{x_i}(t))$ of balls by the Nerve Theorem whenever the balls $B(x_i, r_{x_i}(t))$ are convex for all x_i. Much like an unweighted Čech complex, it takes too much time to compute a weighted Čech complex in practice, so researchers instead usually compute a *weighted VR complex* $\mathrm{VR}_t^{\mathrm{weighted}}(X, M, d, \{w_i\})$. This is the simplicial complex whose vertices are $\{x_i \mid w_i < t\}$ and whose simplices $[x_{i_0}, \ldots, x_{i_k}]$ satisfy $d(x_{i_j}, x_{i_\ell}) + w_{i_j} + w_{i_\ell} < 2t$. The sequence $\left\{\mathrm{VR}_t^{\mathrm{weighted}}(X, M, d, \{w_i\})\right\}_t$ for increasing t is the *weighted VR filtration*. Analogously to (4.1), the weighted VR filtration approximates the weighted Čech filtration in the sense that

$$\check{\mathrm{C}}_r(X, M, d, \{w_i\}) \subseteq \mathrm{VR}_r(X, M, d, \{w_i\}) \subseteq \check{\mathrm{C}}_{\sqrt{2}r}(X, M, d, \{w_i\}) \tag{4.3}$$

for all r.

Given a filtration $\mathcal{K}_{\alpha_0} \subseteq \cdots \subseteq \mathcal{K}_{\alpha_n}$, one can compute the homology of each simplicial complex $\mathcal{K}_{\alpha_i}$. As discussed in Section 4.1, a homology class represents a hole that exists in a filtration for some range of filtration-parameter values α_i. A 0D homology class represents a connected component, and a 1D homology class represents a hole that is bounded by a closed path. To see why 0D homology classes are "holes", we note that one can also view a 0D homology class as representing the empty region between connected components.

As the filtration parameter α_i grows, holes form and subsequently fill in. The information that is given by the birth and death of the homology classes of a filtration is called the *persistent homology* (PH) of the filtration. We say that a homology class is *born* at α_i if i is the minimum index such that the homology class appears in K_{α_i}. Its *birth simplex* is the simplex that creates the homology class. For example, in Figure 4.2, a 1D homology class is born at filtration-parameter value 2. Its birth simplex is the edge with vertices 0 and 3. A homology class that is born at α_i subsequently *dies* at α_j, with $j \geq i$, if j is the minimum index such that the homology class becomes trivial (i.e., the corresponding hole fills in) in K_{α_j}. We refer to α_i as the homology class's *birth value* and to α_j as its *death value*. Its *death simplex* is the simplex that destroys the homology class. For example, the homology class that is born at filtration-parameter value 2 in Figure 4.2 subsequently dies at filtration-parameter value 4. Its death simplex is the triangle with vertices 0, 2, and 3.

One can summarize PH in a *persistence diagram* (PD) [DW22]. A PD is a multiset of points in the extended plane $\mathbb{R}^2 \cup \{\infty\}$. For a homology class with birth value b and death value d, the PD includes a point with coordinates (b, d). An infinite death time d represents a homology class that is not filled in at any filtration $r > b$. In Figure 4.4, we show the PD for the PH of the filtration in Figure 4.2.

In our application of PH to polling-site coverage in Chapter 5, we interpret homology classes as holes in coverage and we interpret the death simplices as the locations of the holes.

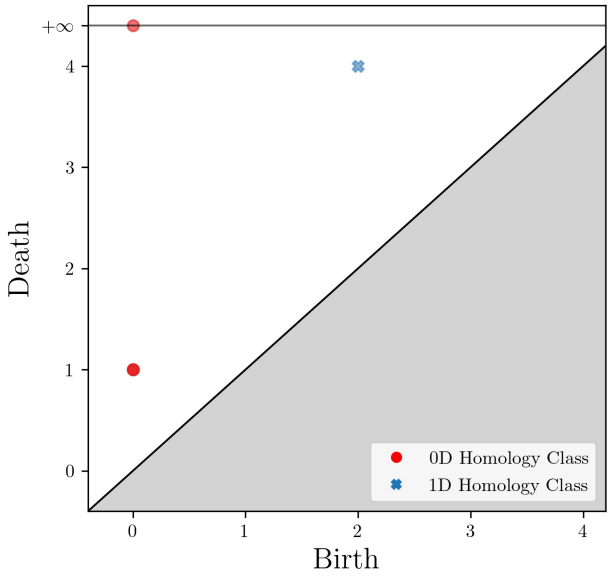


Figure 4.4: The persistence diagram for the 0D and 1D PH of the filtration in Figure 4.2.

CHAPTER 5

Persistent Homology for Resource Coverage: A Case Study of Access to Polling Sites

In this chapter, we use *persistent homology* (PH), which is a tool from *topological data analysis* (TDA), to study the availability and coverage of polling sites. We discussed the background on these topics in Chapter 4. The information from PH allows us to infer "holes" in a distribution of polling sites. We analyze and compare the coverage of polling sites in Los Angeles County and five cities (Atlanta, Chicago, Jacksonville, New York City, and Salt Lake City). This chapter is adapted from [HJJ24], which was led jointly by Abigail Hickok, Benjamin Jarman, Jiajie Luo, and me and was coauthored with Mason A. Porter.[1] All figures in this chapter appeared originally in [HJJ24].

This chapter proceeds as follows. We describe our approach in Section 5.2, present and examine persistence diagrams in Section 5.3, and conclude and discuss implications, limitations, and potential future directions of our work in Section 5.4. Our code is available at `https://bitbucket.org/jerryluo8/coveragetda/src/main/`.

[1] I codesigned the methodology with Abigail Hickok, Benjamin Jarman, and Jiajie Luo; computed the distance matrices (see Section 5.2) with Benjamin Jarman and Jiajie Luo; and wrote the paper with all coauthors.

5.1 Introduction

The geographical distribution of resources such as polling sites (i.e., locations where people vote), hospitals, COVID-19 vaccination sites, Department of Motor Vehicles (DMV) locations, and Planned Parenthood clinics is a major factor in the equitability of access to those resources. Consequently, given the locations of a set of resource sites, it is important to quantify their geographical coverage and to identify underserved geographical regions (i.e., "holes in coverage").

A naive approach to quantifying resource coverage is to consider the geographical distances from resource sites by simply calculating the percentage of people who reside within some cutoff distance D of the nearest resource site. This naive approach is common in policy. For example, in March 2021, United States President Joseph Biden announced a goal to ensure that at least 90% of the adult US population is within 5 miles (i.e., $D = 5$ miles) of a COVID-19 vaccination site [The21]. As another example, it is required by Indian law that 100% of voters live within 2 km of a polling site [SSH19] (i.e., $D = 2$ km). However, such an approach poses at least two issues:

(1) it requires choosing an arbitrary cutoff distance D; and

(2) using only geographical distance fails to account for many other factors, such as population density and the availability (and facility) of public transportation, that affect ease of access to a resource.

These issues severely limit the utility of this naive approach.

In the present chapter, we use PH to study holes in resource coverage. We use PH to analyze data in the form of a *point cloud* $X = \{x_i\}_{i=1}^n$ of points in a metric space (M, d).[2] In this chapter, X is a collection of resource sites, with specified latitudes and longitudes, and

[2]One can weaken the requirement that d is a metric. In this chapter, we use a distance function d that is not technically a metric because it does not satisfy the triangle inequality.

$M = \mathbb{R}^2$ with a non-Euclidean distance function d (see Section 5.2). Given a point cloud X and a scale parameter $r > 0$, one can consider the *r-coverage* $C_r := \bigcup_{i=1}^{n} B(x_i, r)$. As the scale parameter r grows, holes arise and subsequently fill in. PH tracks the formation and disappearance of these holes. When a point cloud is a collection of resource sites, one can interpret holes that persist for a large range of r as holes in coverage. Our approach using PH gives a way to measure and evaluate how equitably a resource is distributed geographically.

Our approach addresses both of the issues (see points (1) and (2)) of the naive approach that we discussed above. First, PH eliminates the need to choose an arbitrary cutoff distance because one can study holes in coverage at all scales. Second, instead of employing geographical distance, we construct a distance function d that is based on travel times. We also incorporate the waiting time at each resource site by constructing a weighted Vietoris–Rips (VR) filtration (see Section 4.2). In a city with a high population density or a poor transportation system, the time that is spent waiting at or traveling to a resource site can be a much higher barrier to access than geographical distance [GS03,HK05]. We estimate waiting times using Global Positioning System (GPS) ping data from mobile phones at the resource sites, and we estimate travel times using street-network data, per capita car-ownership data, and the Google Maps application programming interface (API) [Goo]. Using these estimates, we construct a weighted VR filtration. We weight vertices by our estimates of waiting times, and we define the distance between two vertices to be the estimated round-trip travel time between them. Because the weighted VR filtration is stable, small errors in our estimates cause only small errors in the resultant PH [ACG19].

In this chapter, we examine polling sites as a case study of using PH to study the coverage of resource sites. We restrict our attention to six cities[3]: Atlanta, Chicago, Jacksonville (in Florida), Los Angeles[4], New York City (NYC), and Salt Lake City. We use these cities in part

[3]Although we frame our discussion in terms of cities, some organizations instead use counties when considering the coverage of polling sites.

[4]For Los Angeles, we actually study Los Angeles County. We discuss the reasons for this choice in Section 5.2.5.

because data about them (e.g., car-ownership data) is widely available. Additionally, these cities differ considerably in their demographics and infrastructures, and we can thus compare a variety of different types of cities. Atlanta and New York City are both infamous for long waiting times at polling sites, especially in non-White neighborhoods [Fow20, Kan19]. In 2020, some counties in the Atlanta metropolitan area had a mean of 3600 voters per polling site; the number of polling sites had been cut statewide in Georgia by 10% since 2013 [Fow20]. In New York City, each polling site had a mean of 4173 voters in 2018. As a comparison, in 2004, Los Angeles County and Chicago had only an estimated 1300 and 725 voters per polling site, respectively [Kan19]. However, Los Angeles is infamous for its traffic [Sch21], which can affect voters' travel times to polling sites. Los Angeles and Chicago also differ in the quality of their public transportation, which also affects travel times to polling sites. In our investigation, we seek both to compare the coverage of polling sites in our six focal cities and to identify underserved areas within each city.

5.1.1 Related Work

One can use tools from geography to study resource accessibility. Pearce, Witten, and Bartie [PWB06] used a geographical-information-systems (GIS) approach to examine the accessibility of community resources and how it affects health. Hawthorne and Kwan [HK12] used a GIS approach and a notion of perceived distance to measure healthcare inequality in low-income urban communities. Brabyn and Barnett [BB04] illustrated that there are regional variations in geographical accessibility to general-practitioner doctors in New Zealand and that these regional variations depend on how one measures accessibility.

Another motivation for our study of resource-site coverage is the related problem of sensor coverage. Given a set S of sensors in a domain $\Omega \subseteq \mathbb{R}^2$, one seeks to determine whether every point in Ω is within sensing range of at least one sensor in S. Typically, each sensor has a fixed, uniform sensing radius r_s. In this case, the problem is equivalent to determining whether or not the domain Ω is covered by balls of radius r_s around each $s \in S$. In

[SG06, SG07], de Silva and Ghrist gave homological criteria for sensor coverage. Approaches to studying sensor coverage that use computational geometry (specifically, approaches that involve the Voronoi diagram of S and the Delauney triangulation of S) were discussed in [LWF03, MKP01].

Our problem is also a coverage problem, but there are important differences. The key conceptual difference is that we consider neighborhoods whose sizes depend on a filtration parameter, rather than neighborhoods with a fixed size. Additionally, we do not seek to determine whether or not balls of any particular radius cover a domain; instead, our goal is to quantify the coverage at all choices of radius and to determine how the holes in coverage evolve as we increase the filtration parameter. Another difference between our work and sensor-coverage problems is that our point cloud represents a set of resource sites (in particular, polling sites), rather than a set of sensors. In a sensor network, pairwise communication between sensors can play a role in whether or not the sensors are fully *connected* to each other (in a graph-theoretic sense) and in determining whether or not a domain is covered [ZH05]. By contrast, communication between resource sites does not play a role in access to those resource sites.

Several studies include applications of PH to geospatial data [CJ23, Fen25]. Feng and Porter [FP21] developed two methods to construct filtrations—one that uses adjacency structures and one that uses the level-set method [OF03] of front propagation—and applied their approaches to examine geospatial distributions of voting results in the 2016 United States presidential election. They identified *political islands* (i.e., precincts that voted more heavily for a candidate than their surrounding precincts). In [FP20], Feng and Porter used their PH approaches to study spatial networks. Friesen and Ziegelmeier [FZ24] used the level-set PH method of [FP21] to examine the structure of racial segregation in US cities. Stolz, Harrington, and Porter [SHP16] used a conventional PH approach to examine the geospatial distribution of voting results in the United Kingdom's "Brexit" referendum. Hickok, Needell, and Porter [HNP22] used PH to study geospatial anomalies in COVID-19 case-rate data (see

also [FHP22]) and vaccination-rate data. Corcoran and Jones [CJ23] used PH to perform (1) a point-pattern analysis of pubs for many UK cities and (2) a spatiotemporal analysis of rainfall in the UK. Kauba and Weighill [KW24] used PH to examine demographic patterns in the Black and Hispanic populations of 100 US cities. Kadeethum and Downs [KD24] used PH to identify undocumented oil and gas wells from satellite images, and O'Neil and Tymochko [OT24] used PH to study holes in cooling-center coverage in US cities.

5.2 Our Construction of Weighted VR Complexes

For each city, we construct a weighted VR filtration in which the point cloud $X = \{x_i\}$ is the set of polling sites in $\mathbb{R}^2$ and the weight w_i of a point x_i is an estimate of the waiting time at the corresponding polling site. Instead of computing a weighted VR filtration with respect to Euclidean distance, we define a distance function that estimates the mean amount of time that it takes to travel to and from a polling site. With respect to this distance function, the union $\bigcup_i B(x_i, r_{x_i}(t))$ (see (4.2)) is the set of points y such that the estimated time for an individual at y to vote (including waiting time and travel time[5] in both directions) is at most t. The weighted Čech complex $\check{\mathrm{C}}_t^{\mathrm{weighted}}(X, \mathbb{R}^2, d, \{w_i\})$ is an approximation of $\bigcup_i B(x_i, r_{x_i}(t))$. When the balls $B(x_i, r_{x_i}(t))$ are convex, the weighted Čech complex is homotopy-equivalent to $\bigcup_i B(x_i, r_{x_i}(t))$, so these two complexes have the same homology (i.e., the same set of holes). The weighted VR complex $\mathrm{VR}_t^{\mathrm{weighted}}(X, \mathbb{R}^2, d, \{w_i\})$ is an approximation of the weighted Čech complex.

We construct our distance function as follows. Let x and y be two polling sites. We

[5]Incorporating information (such as waiting times) other than travel times is sensible both in principle and in practice. In our computational experiments, using only travel times yields results that differ drastically from those that we present in this chapter.

estimate the expected time for an individual to travel from x to y and back to be

$$\begin{aligned}\tilde{d}(x,y) := C(Z(x)) \min\left\{t_{\text{car}}(x,y), t_{\text{pub}}(x,y), t_{\text{walk}}(x,y)\right\} \\ + \left[1 - C(Z(x))\right] \min\left\{t_{\text{pub}}(x,y), t_{\text{walk}}(x,y)\right\} ,\end{aligned}$$

where $Z(x)$ is the zip code that includes x (a polling site), $C(Z(x))$ is an estimate of the fraction of voting-age people in $Z(x)$ who can travel by car to a polling site, and $t_{\text{car}}(x,y)$, $t_{\text{pub}}(x,y)$, and $t_{\text{walk}}(x,y)$ are estimates of the expected travel times from x to y and back by car, public transportation, and walking, respectively. We calculate $C(Z(x))$ by dividing an estimate of the number of personal vehicles in $Z(x)$ by an estimate of the voting-age population in $Z(x)$; see Section 5.2.3. We discuss how we calculate t_{car}, t_{pub}, and t_{walk} in Section 5.2.1.

Our definition of $\tilde{d}(x,y)$ captures the cost (in time) to travel to vote. In particular, $\tilde{d}(x,y)$ is an estimate of the mean travel time for an individual who resides in zip code $Z(x)$ to travel from x to y and back. We assume that all individuals choose the fastest mode of transportation that is available to them. Therefore, individuals who can travel by car choose the fastest option between driving, taking public transportation, and walking. Their travel time is $\min\left\{t_{\text{car}}(x,y), t_{\text{pub}}(x,y), t_{\text{walk}}(x,y)\right\}$. Likewise, we assume that individuals who do not have access to a car choose the fastest option between taking public transportation and walking. Their travel time is $\min\left\{t_{\text{pub}}(x,y), t_{\text{walk}}(x,y)\right\}$. Our estimate of the fraction of a population with access to a car is $C(Z(x))$, so the fraction without a car is $1 - C(Z(x))$. Therefore, $\tilde{d}(x,y)$ is the (estimated) mean time for an individual who resides in zip code $Z(x)$ to travel from x to y and back.

The function $\tilde{d}(x,y)$ is not symmetric (i.e., $\tilde{d}(x,y) \neq \tilde{d}(y,x)$) because $C(Z(x)) \neq C(Z(y))$. However, we need a symmetric function to construct a weighted VR filtration. To construct a symmetric distance function that is based on $\tilde{d}(x,y)$, we define the distance between x and y to be a weighted average of $\tilde{d}(x,y)$ and $\tilde{d}(y,x)$, where we determine the weights from the populations of the zip codes that include x and y. More precisely, we define the distance

between x and y to be

$$d(x,y) := \frac{1}{P}[P_{Z(x)}\tilde{d}(x,y) + P_{Z(y)}\tilde{d}(y,x)]\,, \tag{5.1}$$

where $P_{Z(x)}$ and $P_{Z(y)}$ are the populations of zip codes $Z(x)$ and $Z(y)$, respectively, and $P := P_{Z(x)} + P_{Z(y)}$ is the sum of the populations of $Z(x)$ and $Z(y)$. With respect to this distance function, the ball $B(x,r)$ is the set of points y such that the expected time for an individual to travel back and forth between x and y is at most r, where the individual starts randomly at x or y with probabilities that are weighted by the populations of their associated zip codes. Although our distance function is not technically a metric (because it does not satisfy the triangle inequality), we can still construct a weighted VR filtration using the definition in Section 4.2.

5.2.1 Estimating Travel Times

To compute our distance function (see (5.1)), we need to estimate the pairwise travel times by car, public transportation, and walking between each pair of polling sites. We measure these times in minutes.

We estimate the time that it takes to walk between each pair of polling sites using street networks, which are available through the OpenStreetMap tool [Ope], for each of our cities. Using OpenStreetMap, we calculate a shortest path (by geographical distance) between each pair of polling sites. In Figure 5.1, we show an example of a shortest path between two polling sites in Atlanta.

Let $L(x,y)$ denote the length (which we measure in meters) of a shortest path (by geographical distance) between polling sites x and y. Our estimate of the walking time (in minutes) from x to y and back is $t_{\text{walk}}(x,y) := 2L(x,y)/v_{\text{walk}}$, where $v_{\text{walk}} = 85.2$ meters per minute is an estimate of the mean walking speed of an adult human [BBH06].

To estimate the travel times by car and public transportation, we use the Google Maps Distance Matrix API [Goo]. Because of budgetary constraints (and the cost of five dollars

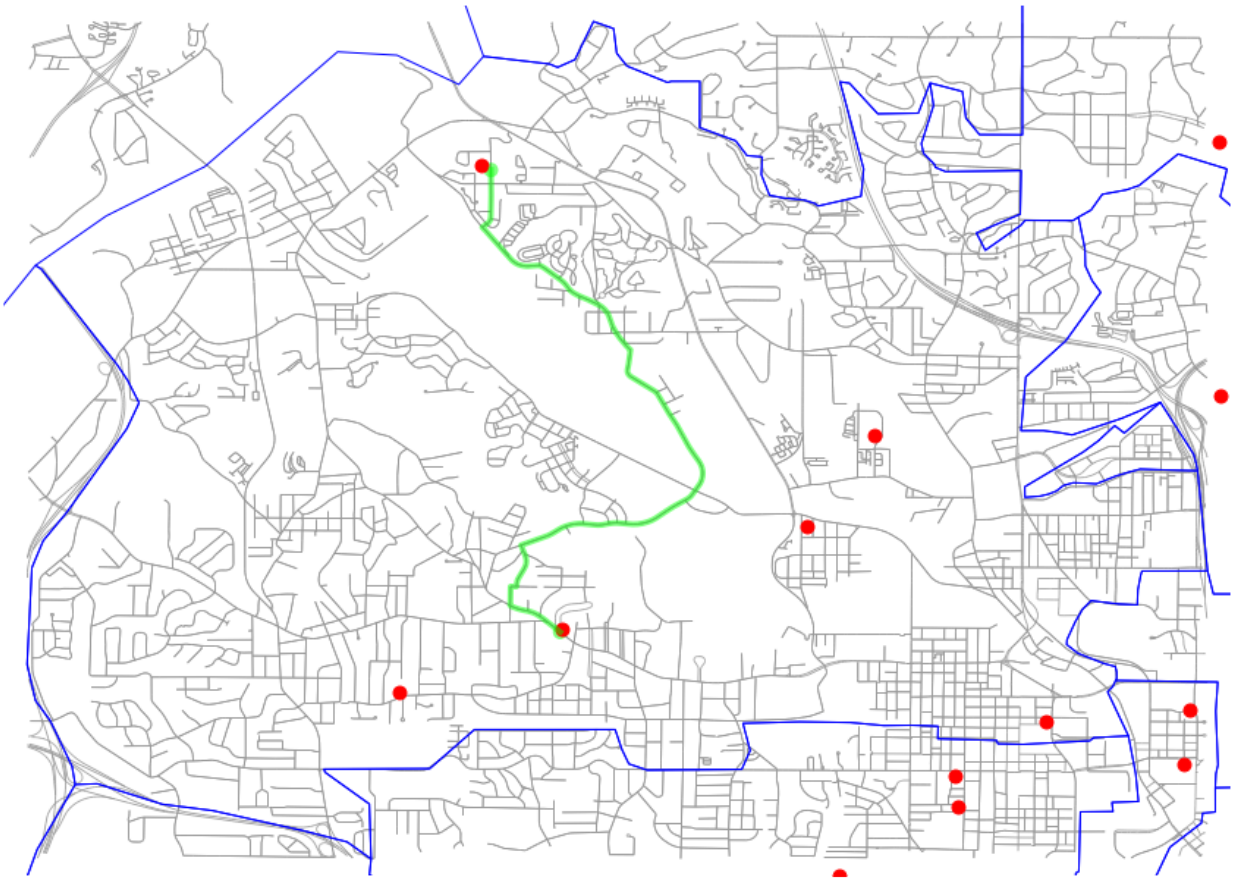


Figure 5.1: A shortest path (by geographical distance) between two polling sites in zip code 30314 in Atlanta.

per thousand API queries), we use this API to estimate only the travel times between each polling site and its 25 geographically closest polling sites. We refer to these sites as a polling site's 25 nearest neighbors.

For each of the 25 nearest neighbors, we separately calculate both the time from a polling site to each neighbor and the time to a polling site from each neighbor. These two travel times are often different because of different traffic conditions or other factors. We estimate the remaining pairwise travel times as follows. Let G be the unweighted, undirected graph whose vertices are the polling sites and whose edges connect each vertex to its 25 nearest

neighbors.[6] Let G_{car} and G_{pub} be the weighted, directed graphs whose vertices and edges[7] are those of G and whose weights are the travel times (by car and public transportation, respectively) that we compute using the Google Maps API. The weight of the directed edge from vertex x to vertex y is the travel time from x to y. Therefore, the weight of the edge from x to y may differ from the weight of the edge from y to x. For any two polling sites x and y, let the travel times $\tilde{t}_{\text{car}}(x, y)$ and $\tilde{t}_{\text{pub}}(x, y)$ be the length of a shortest weighted path from x to y in the graphs G_{car} and G_{pub}, respectively. The corresponding symmetrized travel times $t_{\text{car}}(x, y)$ and $t_{\text{pub}}(x, y)$ are

$$t_{\text{car}}(x, y) := \tilde{t}_{\text{car}}(x, y) + \tilde{t}_{\text{car}}(y, x)\,,$$
$$t_{\text{pub}}(x, y) := \tilde{t}_{\text{pub}}(x, y) + \tilde{t}_{\text{pub}}(y, x)\,.$$

5.2.2 Estimating Waiting Times

Our weighted VR filtrations have weights at each vertex (i.e., polling site) that are given by an estimate of the mean time that a voter spends (i.e., the mean waiting time) at that polling site. In a nationwide study of waiting times at polling sites during the 2016 US presidential election [CHP19], Chen et al. used smartphone data of hundreds of thousands of voters to estimate waiting times. They also examined potential relationships between waiting times and racial demographics.

We construct our waiting-time estimates using the congressional-district-level estimates in [CHP19] (see their Table C.2). For each polling site x, we calculate the mean of the waiting-time estimates for each congressional district that overlaps with the zip code $Z(x)$ that contains x. This averaging procedure yields estimates of waiting times at the zip-code level. (We transform our waiting-time data to the zip-code level because the rest of our data

[6] The relation of being one of a vertex's 25 nearest neighbors is not symmetric. Therefore, the degrees of some vertices are larger than 25.

[7] We view each undirected edge (x_i, x_j) of G as a bidirectional edge, and we include both of the associated directed edges in the directed graphs G_{car} and G_{pub}.

is at the zip-code level.)

5.2.3 Estimates of Demographic Information

We obtain estimates of demographic data at the zip-code level from 2019 five-year American Community Survey data [US]. We use voting-age population data from their Table ACSDT5Y2019.B29001 and vehicle-access data from their Table ACSDT5Y2019.B25046.

5.2.4 Polling-Site Zip Codes

Much of our data is at the zip-code level, and we treat a polling site's zip code as representative of its local area. Certain polling sites (predominantly government buildings) have their own zip codes, despite their populations of 0. We adjust the zip codes of such polling sites to match the zip codes of the directly surrounding areas.

5.2.5 Special Treatments of Our Cities

The city of Atlanta does not include the suburbs of the Atlanta metropolitan area, so we use the entire area that is served by the Atlanta Regional Commission.

Chicago's boundary is not convex (especially in the northwest), so we include all areas of all zip codes, even when only a small portion of a zip code lies within the city of Chicago.

Because of the oddly shaped city boundaries of Los Angeles, which surrounds several exclaves, we use the entirety of Los Angeles County (except for its islands).

Because of the disconnected nature of New York City, we subdivide it into three regions (Queens and Brooklyn, Manhattan and the Bronx, and Staten Island) and treat each region separately. We then combine our results for the three regions into a single presentation. For example, we combine the PDs into a single PD for all of New York City.

For more information about the treatment of each city, see the file "readme.txt" in our

repository at `https://bitbucket.org/jerryluo8/coveragetda/src/main/`.

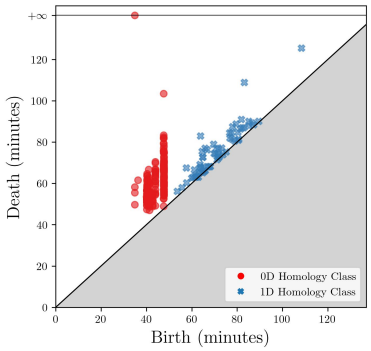
5.3 Results

We compute the PH of the weighted VR filtrations of Section 5.2 for Atlanta, Chicago, Jacksonville, Los Angeles County, New York City, and Salt Lake City. We show their PDs in Figure 5.2. We examine 0D and 1D homology classes. The 0D homology classes represent holes between different connected regions of coverage, and the 1D homology classes represent holes in coverage that are bounded by closed paths. A homology class that dies at filtration-parameter value t represents a hole in coverage that persists until time t. An individual who lives in a hole in coverage that dies at t needs t minutes (including both waiting time at a polling site and travel time back and forth to the site) to cast a vote.

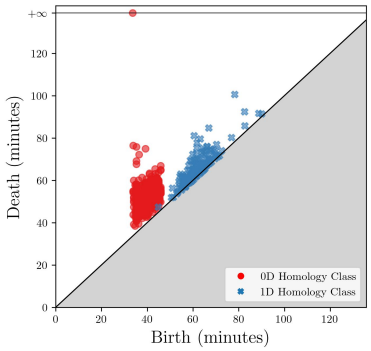
In our analysis, we emphasize homology-class death values. We view homology-class birth values as largely irrelevant to our application. A homology-class birth value indicates the filtration-parameter value at which a coverage hole materializes. We use birth values only in the following way. If the death value divided by the birth value (i.e., the *death/birth ratio*) of a homology class is very small (i.e., it is close to 1), then it is possible that this class is merely an artifact of using a VR approximation of a Čech complex. We thus focus on homology classes whose death/birth ratios are at least 1.05.[8] Beyond this, we use only the homology-class death values and death simplices.

Larger homology-class death values suggest that a city may have worse coverage, and a wider distribution of death values suggests that there may be more variation in polling-site accessibility within a city. In Figure 5.3, we show a box plot of the distribution of homology-class death values for each city. In Table 5.1, we show the medians and variances of the 0D and 1D homology-class death values for each city.

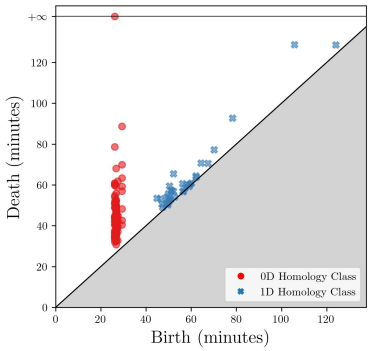
[8]Interested readers can explore thresholds other than 1.05 using our data, which is available at `https://bitbucket.org/jerryluo8/coveragetda/src/main/`. We describe the data in detail in the file "readme.txt".



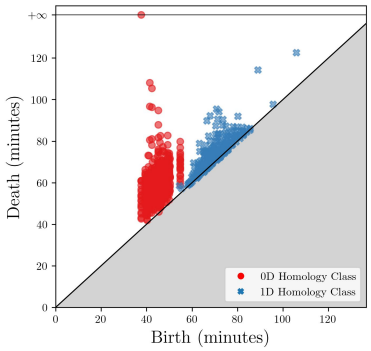
(a) Atlanta



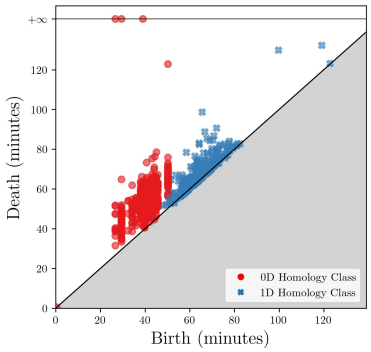
(b) Chicago



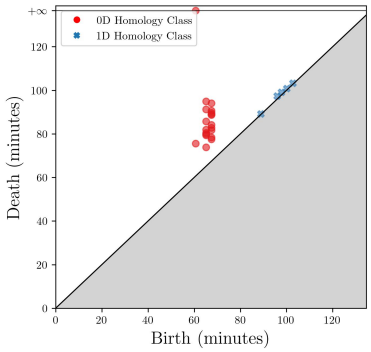
(c) Jacksonville (Florida)



(d) Los Angeles County



(e) New York City



(f) Salt Lake City

Figure 5.2: The PDs for each city for the PH of our weighted VR complexes.

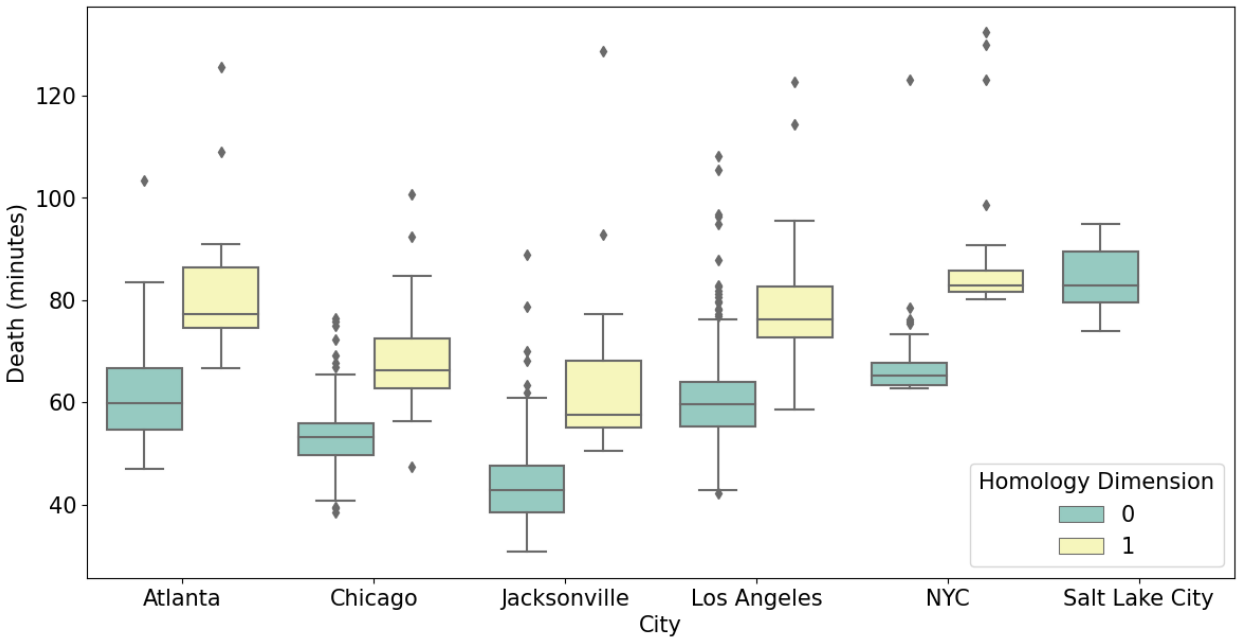
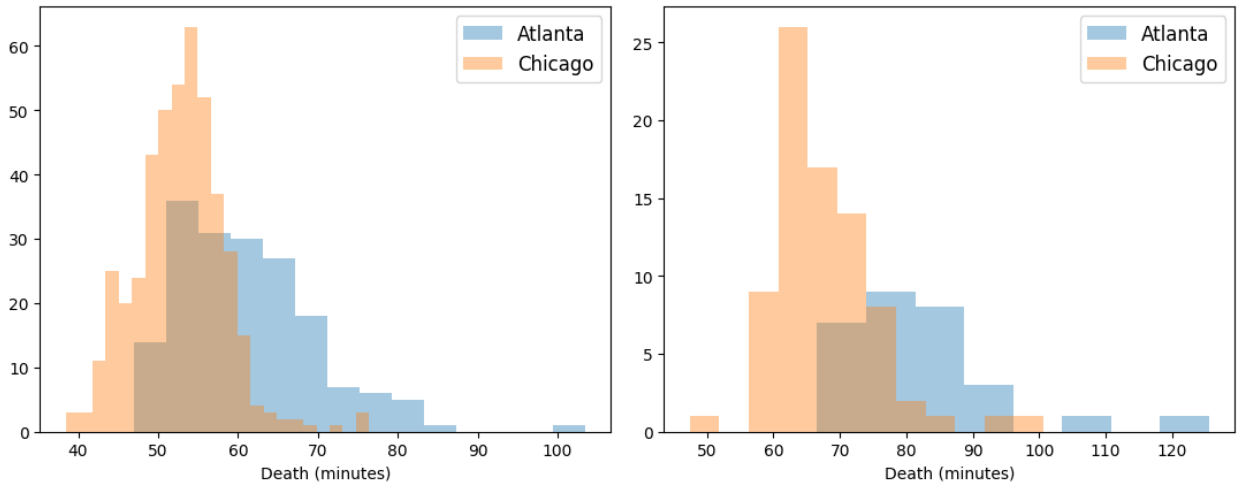


Figure 5.3: Box plots of the death values of the 0D and 1D homology classes for each city. We only consider homology classes whose death/birth ratio is at least 1.05. Salt Lake City has no such 1D homology classes.

City	Homology Dimension	Median (minutes)	Variance (minutes)
Atlanta	0	59.9	75.4
	1	77.1	150.8
Chicago	0	53.1	30.2
	1	66.3	59.7
Jacksonville (Florida)	0	42.8	75.7
	1	57.5	394.4
Los Angeles County	0	59.6	53.3
	1	76.1	84.6
New York City	0	65.1	49.2
	1	82.9	207.1
Salt Lake City	0	82.8	37.3
	1	N/A	N/A

Table 5.1: The medians and variances of the homology-class death values for each city. (As we discussed in the main text, we consider Los Angeles County rather than only the city of Los Angeles.) We consider homology classes whose death/birth ratio is at least 1.05. Salt Lake City has no such 1D homology classes.



(a) 0D homology classes

(b) 1D homology classes

Figure 5.4: Histograms of the death values of the 0D and 1D homology classes for Atlanta and Chicago. We only consider homology classes whose death/birth ratio is at least 1.05.

We compare the coverages of the cities by examining the death values in the PDs. For example, in the PDs for Atlanta and Chicago in Figure 5.2, we see that Atlanta's homology classes tend to die later than Chicago's homology classes. We also see this in the box plots in Figures 5.3 and 5.4, in which we plot the distributions of death values for Atlanta and Chicago. Our PDs and visualizations of summary statistics suggest that Chicago has better polling-site coverage than Atlanta.

We use the death simplices to locate and visualize holes in polling-site coverage. We interpret the death simplex of a homology class as the *epicenter* of an associated coverage hole because the death simplex represents the last part of the hole to be covered. The death simplex of a 0D homology class is an edge between two polling sites; there is a hole in coverage between those two sites. Similarly, the death simplex of a 1D homology class is a triangle that is the convex hull of three polling sites; there is a hole in coverage between those three sites. In Figures 5.5 and 5.6, we show the death simplices with the largest death

values[9] for the 0D and 1D homology classes,[10] respectively. For example, consider panels (a) and (b) of Figures 5.5 and 5.6, in which we show the death simplices of the 0D and 1D homology-classes for Atlanta and Chicago. The areas of lowest coverage (i.e., the areas that have the death simplices with the largest death values) in Atlanta tend to be in the southwest, whereas the areas of lowest coverage in Chicago tend to be in the northwest and southeast. There is one 1D homology class in Atlanta that has a significantly larger death filtration value than the other classes in Atlanta and any of the classes in Chicago. This homology class represents a hole in coverage in southwest Atlanta (see Figure 5.6a).

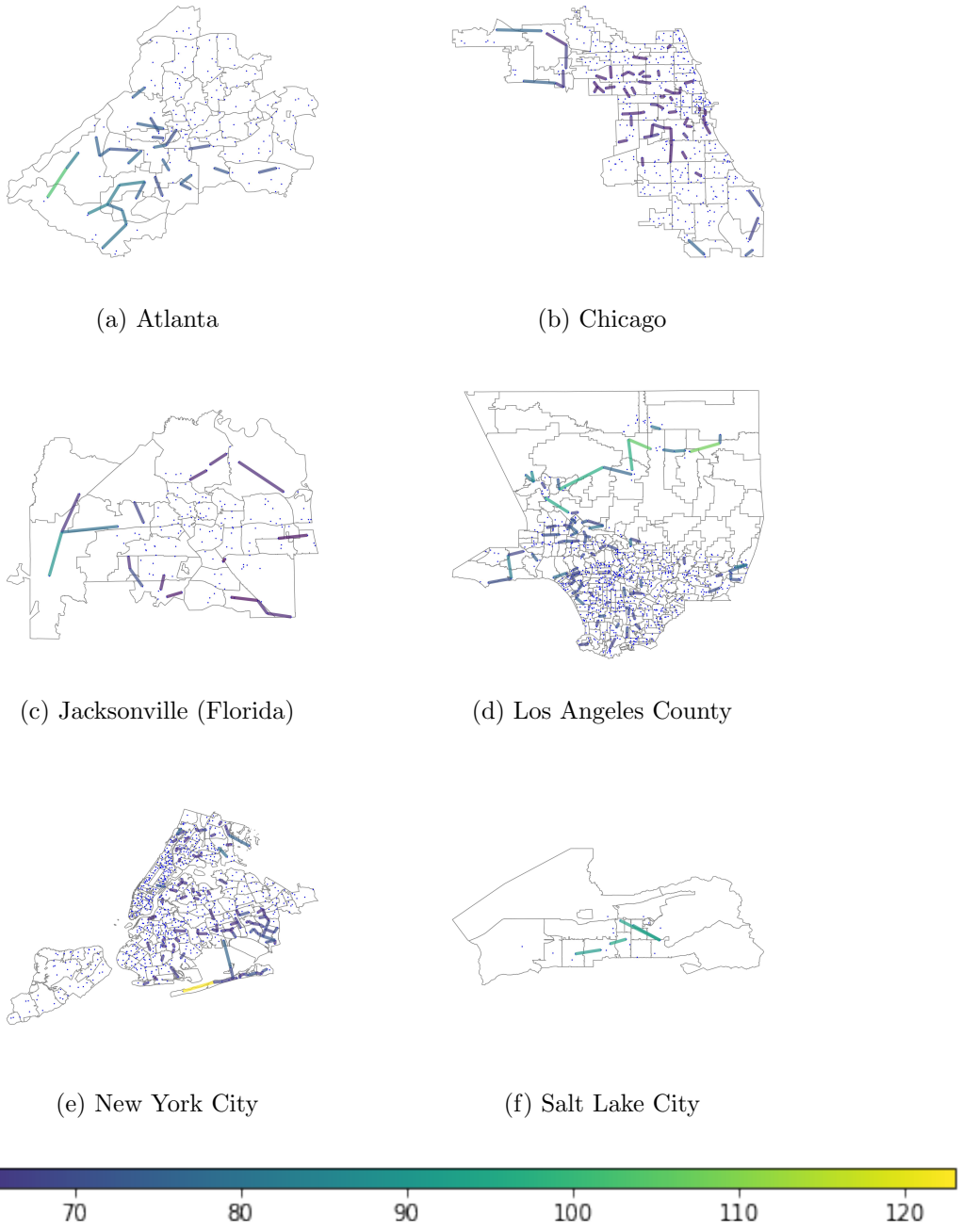
5.4 Conclusions and Discussion

5.4.1 Summary

We showed that PH is a helpful approach to studying accessibility and equitability of resources. It allows one to examine holes in resource coverage with respect to an appropriate choice of *distance*, which one constructs to incorporate important features of a problem of interest. The distance can be based on geography, time, or something else. In the present chapter, we used PH to study and quantify holes in polling-site coverage in six US cities (technically, in five cities and Los Angeles County). For each city, we constructed a filtration in which a homology class that dies at time t represents a geographical region in which it

[9]More precisely, for each city and each homology dimension (0 and 1), we show the death simplices whose death values have a z-score of at least 1. We calculate the z-score as follows. Let d be the death value of a p-dimensional homology class (where $p = 0$ or $p = 1$) for city C. The z-score of d is $z = (d - \mu_{C,p})/\sigma_{C,p}$, where $\mu_{C,p}$ and $\sigma_{C,p}$ are the mean and standard deviation of the distribution of death values of the p-dimensional homology class for city C.

[10]In Figure 5.6, in which we show the death simplices of the 1D homology classes, some of the polling sites appear to be covered by death simplices whose vertices are other polling sites. At least two factors may contribute to this. One factor is that our measure of distance is not a Euclidean metric, even though we plot the death simplices in Figure 5.6 as Euclidean triangles. The Euclidean triangles can sometimes cover polling sites that are not among its vertices, but geodesic triangles may not cover those polling sites. Another possibility is that a polling site x has such a long waiting time that it does not show up in the filtration until after the homology class whose death simplex includes x has already died.



(a) Atlanta

(b) Chicago

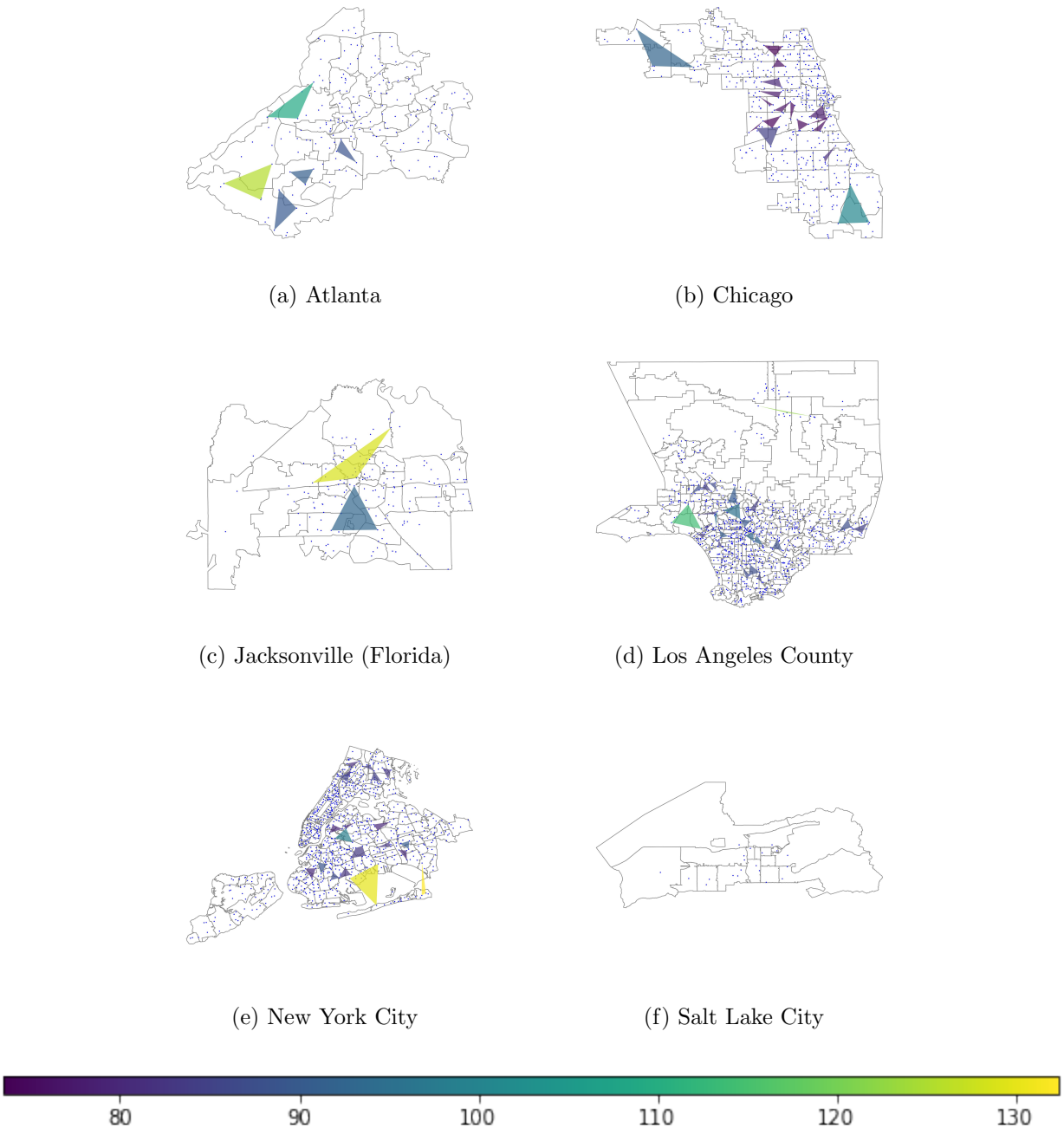
(c) Jacksonville (Florida)

(d) Los Angeles County

(e) New York City

(f) Salt Lake City

Figure 5.5: Death simplices with the largest death values for the 0D homology classes. The colors correspond to the death values (in minutes). We only consider homology classes whose death/birth ratio is at least 1.05.



(a) Atlanta

(b) Chicago

(c) Jacksonville (Florida)

(d) Los Angeles County

(e) New York City

(f) Salt Lake City

Figure 5.6: Death simplices with the largest death values for the 1D homology classes. The colors correspond to the death values (in minutes). We only consider homology classes whose death/birth ratio is at least 1.05.

takes t minutes to cast a vote (including both travel time and waiting time). We interpreted the death simplex of a homology class as the location of the corresponding hole in resource coverage. The information in the PH allowed us both to compare the accessibility of voting across our chosen cities and to determine the locations of the coverage holes within each city.

A key benefit of our use of PH is that it enabled us to identify holes in polling-site coverage at all time scales. It also allowed us to use a distance that we designed for the problem at hand, rather than merely using geographical distance, which does not capture important factors in resource accessibility [BB06]. We based our distance function on estimates of travel time, which is more reasonable and accurate than geographical distance for capturing resource accessibility [PWB06].

5.4.2 Limitations

To conduct our study, we needed to estimate a variety of quantities (see Section 5.2), including travel times, waiting times, and demographic information. We also made several simplifications because of computational and monetary constraints. We now discuss some issues that are important to address before attempting to incorporate our approach into policy-making.

One limitation of our study is our estimation of travel times. As we discussed in Section 5.2.1, we computed travel times using the Google Maps API. Because of monetary constraints, we only computed a subset of the relevant travel times and used a graph-based estimate to determine the others. Additionally, we computed each travel time between polling sites only once. Computing more precise estimates of travel times is important to better capture the accessibility of polling sites. One way to do this is to compute travel times between the same two polling sites multiple times across different days and times of day and take an average. Such additional computations can also help yield estimates of best-case and worst-case scenarios.

Another limitation of our study is the granularity of our data. As we discussed in Section 5.2.2, our waiting-time data is at the scale of congressional districts. Because there is heterogeneity in the waiting times at different polling sites in the same congressional district, it is important to obtain finer-grained data for the waiting times at polling sites. Having finer-grained waiting times (e.g., if possible, procuring an estimated waiting time for each polling site) would improve our ability to capture voting accessibility.

We also made several topological approximations. We worked with a weighted VR filtration, which approximates a weighted Čech filtration, which in turn approximates the nested set $\{\bigcup_i B(x_i, r_{x_i}(t))\}_{t\in\mathbb{R}}$ of spaces, where $\{x_i\}$ is a set of polling sites and $r_{x_i}(t)$ is the radius function that we defined in Section 5.2. The nested set of spaces is directly relevant to our application, as the holes in $\bigcup_i B(x_i, r_{x_i}(t))$ are the true holes in polling-site coverage. We made our approximations, which are standard in TDA and are well-justified (see our discussion in Section 4.2) [OPT17], to reduce computational cost. However, the convexity condition of the Nerve Theorem, which justifies the approximation of $\bigcup_i B(x_i, r_{x_i}(t))$ by a weighted Čech complex, is not guaranteed to be satisfied for all times t. The Nerve Theorem implies that the weighted Čech complex is homotopy-equivalent to $\bigcup_i B(x_i, r_{x_i}(t))$ whenever the balls $B(x_i, r_{x_i}(t))$ are convex. This condition always holds in Euclidean space, but it is not guaranteed to hold in the space that we defined in Section 5.2.[11] Homotopy-equivalence is important because homotopy-equivalent spaces have the same homology and thus have the same set of holes.

Finally, our approach only detects holes in the convex hull of a set of resource sites.

[11] Although our space is not Euclidean, it is still reasonable to assume that it is approximately *locally* Euclidean. That is, for each polling site x, there is a constant $a > 0$ such that if y is sufficiently close to x, then $d(x, y) \approx a \cdot d_E(x, y)$, where $d(x, y)$ is defined by (5.1) and $d_E(x, y)$ is the Euclidean distance. This approximation holds because car-ownership rates and traffic conditions do not vary much within a sufficiently small neighborhood. We verified empirically that our distance function is approximately locally Euclidean by showing that, for each polling site x, there is a strong linear correlation between the pairwise distances $d(x, y)$ and the pairwise Euclidean distances $d_E(x, y)$ when y is sufficiently close to x. Because our distance function is approximately locally Euclidean, sufficiently small balls (with respect to our distance function) behave like Euclidean balls, so the Nerve Theorem is applicable for sufficiently small filtration values.

Although this may be inconsequential if resource sites are sufficiently spread out geographically, it can be problematic if the resource sites are overly concentrated near a few locations. One way to address this issue is to incorporate city boundaries into the construction of the filtrations. This would help capture holes in coverage in regions that lie outside the convex hull of the resource sites, and it would also help identify the filtration-parameter value t at which an entire city is covered by the balls $B(x_i, r_{x_i}(t))$.

5.4.3 Future Work

As we discussed in Section 5.4.2, we made several topological approximations of our mathematical object of interest, which is the nested set $\{\bigcup_i B(x_i, r_{x_i}(t))\}_{t\in\mathbb{R}}$ of spaces. Instead of using a weighted VR filtration, one can construct a more direct approximation of $\{\bigcup_i B(x_i, r_{x_i}(t))\}_{t\in\mathbb{R}}$. One can first discretize a city by imposing a grid onto it. For each point on such a grid, one can then construct the filtered cubical complex that is induced by the travel time to the nearest polling site. However, this is much more computationally expensive than our approach, and it would also entail many more travel-time queries (which cost money) than in our work.[12]

It is also important to incorporate city boundaries into the construction of filtrations. One way to do this is as follows. Let $x_1, \ldots, x_n$ denote the resource sites, and let $y_1, \ldots, y_m$ denote the points that one obtains by discretizing a city boundary. One can extend our distance function (5.1) by defining[13]

$$d(x_i, y_j) := \frac{2}{P}[P_{Z(x_i)}\tilde{d}(x_i, y_j) + P_{Z(y_j)}\tilde{d}(y_j, x_i)]\,, \tag{5.2}$$

[12]Our distance function (5.1) is symmetric, but recall that it is not a metric because it does not satisfy the triangle inequality. Therefore, we cannot use techniques such as distance transforms and level-set propagation to reduce the computational complexity of calculating the filtration $\{\bigcup_i B(x_i, r_{x_i}(t))\}_{t\in\mathbb{R}}$.

[13]The factor of 2 arises from the fact that x_i is a resource site but y_j is not.

where P, P_Z, and $\tilde{d}$ are as in the distance function (5.1) and

$$d(y_i, y_j) = \begin{cases} 0\,, & y_i \text{ and } y_j \text{ are adjacent points of the discretized city boundary} \\ \infty\,, & \text{otherwise}\,. \end{cases} \tag{5.3}$$

At each filtration-parameter value, the simplicial complex that one constructs using the distance function (5.1) with the extensions (5.2) and (5.3) includes both the points that one obtains by discretizing the boundary and the edges that connect adjacent boundary points. The largest death value is then the filtration-parameter value t that corresponds to the time at which an entire city is covered by the balls $\{B(x_i, r_{x_i}(t))\}$ (i.e., when there are no longer any holes in coverage).

We used death simplices to locate holes in coverage, but other approaches are also possible. For example, by calculating minimal generators [LTH21], one can identify representative cycles that encircle holes. The topological pipeline "hyperTDA" was introduced recently [BYM22] to analyze the structure of minimal generators by constructing a hypergraph, calculating hypergraph centrality measures, and employing community detection. This approach may provide insights into the spatial structure of minimal generators. Another potentially viable approach is to use decorated merge trees (DMTs) [CHM22] to locate holes in coverage. DMTs allow one to match holes with associated clusters of points.

Although we have explored a specific case study (namely, the accessibility of polling sites), it is also relevant to conduct similar investigations for other resources, such as public parks, hospitals, vaccine distribution centers, grocery stores, Planned Parenthood clinics, and Department of Motor Vehicles (DMV) locations. One can use similar data to construct a filtration, although it may be necessary to modify the choices of distance and weighting. One can also use ideas from mobility theory [BBG18] to help construct suitable distances and weightings. For example, all DMV offices offer largely the same services, so it seems reasonable to assume that people will go to their nearest office. Therefore, in a study of DMV accessibility, it seems appropriate to use travel time as a distance function, just as

we did in our analysis of polling sites. However, in other applications, it is not reasonable to use travel time alone as a distance function. For example, different grocery stores[14] may offer different products at different prices, so travel time alone may not be appropriate as a choice of distance function. Additionally, although waiting time is a significant factor for investigating the coverage of polling sites, there are many applications for which it does not make sense to incorporate waiting time. For example, the time that is spent in a public park or recreation center is typically not a barrier to access. In applications in which waiting times are not an accessibility factor, it seems more appropriate to use a standard VR filtration than a weighted VR filtration. With salient modifications (such as those that we described in this subsection and in Section 5.4.2), one can apply our approach to many other types of resource sites.

[14]See [HLS24] for a recent study of grocery-store accessibility.

CHAPTER 6

Conclusion

In this dissertation, we presented two projects in network science and relevant background on these topics. In Chapter 1, we introduced the topics in this dissertation. In Chapter 2, we provided background on fundamental ideas in network science. In Chapter 3, we presented and analyzed a novel modeling framework for interacting hosts with microbiome exchange. In Chapter 4, we discussed relevant background on persistent homology. In Chapter 5, we developed our PH methodology for the classification of resource coverage and examined the coverage of polling sites in the 2016 US presidential election.

6.1 Interacting Hosts with Microbiome Exchange

In Chapter 3, we developed a novel framework to model the microbiome dynamics of living hosts that incorporates both the local dynamics within an environment and exchanges of microbiomes between environments. Our framework extends existing metacommunity theory by accounting for the discrete nature of host interactions. Unlike classical mass-effects models, our framework incorporates two distinct parameters that control interaction frequencies and interaction strengths. Using both analytical approximations and numerical computations, we demonstrated that both parameters are necessary to determine microbiome dynamics.

We developed approximations in three parameter regions, and we proved their accuracy in those regions. Our low-frequency approximation (LFA) gives a good approximation of mi-

crobiome dynamics when the local dynamics are much faster than the interactions between hosts. Our high-frequency, low-strength approximation (HFLSA) gives a good approximation when interactions are frequent but weak, resulting in a model with the same form as the mass-effects model (3.3). Finally, our high-frequency, constant-strength approximation (HFCSA) accurately predicts the rapid convergence of all hosts' microbiome dynamics when interactions are frequent and have constant interaction strength. We validated the three approximations through numerical experiments on an illustrative model of microbiome dynamics for a range of parameter values.

Our modeling framework provides a foundation for many promising future research directions in microbiome dynamics. In our framework's current form, one can use it to study the effects of host interactions in many ecological models of local dynamics. One can also use it to study the impact of the structure of interaction networks on microbiome dynamics. There are many possible extensions of our modeling framework. We discuss several extensions in detail in Section 3.6.2.

6.2 Persistent Homology for Resource Coverage

In Chapter 5, we showed that persistent homology (PH), which is a type of topological data analysis (TDA), is a helpful approach to studying accessibility and equitability of resources. It allows one to examine holes in resource coverage with respect to an appropriate choice of distance, which one constructs to incorporate important features of a problem of interest. The distance can be based on geography, time, or something else. We used PH to study and quantify holes in polling-site coverage in six US cities (technically, in five cities and Los Angeles County). For each city, we constructed a filtration in which a homology class that dies at time t represents a geographical region in which it takes t minutes to cast a vote (including both travel time and waiting time). We interpreted the death simplex of a homology class as the location of the corresponding hole in resource coverage. The

information in the PH allowed us both to compare the accessibility of voting across our chosen cities and to determine the locations of the coverage holes within each city.

A key benefit of our use of PH is that it enabled us to identify holes in polling-site coverage at all time scales. It also allowed us to use a distance that we designed for the problem at hand, rather than merely using geographical distance, which does not capture important factors in resource accessibility. We based our distance function on estimates of travel time, which is more reasonable and accurate than geographical distance for capturing resource accessibility.

Our method had a variety of limitations, most of which stemmed from inadequate access to data. We proposed techniques for improving the accuracy of our method when enough data and computational resources are available. It is also important to incorporate city boundaries into the construction of filtrations. This allows classification of polling-site coverage outside the convex hull of polling sites.

Although we explored a specific case study (namely, the accessibility of polling sites), it is also relevant to conduct similar investigations for other resources, such as public parks, hospitals, vaccine distribution centers, grocery stores, Planned Parenthood clinics, and Department of Motor Vehicles (DMV) locations. One can use similar data to construct filtrations, although it may be necessary to modify the choices of distance and weighting. For example, different grocery stores may offer different products at different prices, so travel time alone may not be appropriate as a choice of distance function.

6.3 Final Thoughts

In this dissertation, we presented two projects in network science. In the first project, we took a theory-driven approach to studying the microbiomes of interacting hosts. We developed and analyzed a novel modeling framework that captures the discrete nature of host interactions. In the second project, we applied ideas from algebraic topology to data

analysis. We used persistent homology to classify the coverage of resource sites. These projects employed different mathematical approaches, yielding distinct insights into their respective applications.

REFERENCES

[ACG19] Hirokazu Anai, Frédéric Chazal, Marc Glisse, Yuichi Ike, Hiroya Inakoshi, Raphaël Tinarrage, and Yuhei Umeda. "DTM-Based Filtrations." In Gill Barequet and Yusu Wang, editors, *35th International Symposium on Computational Geometry (SoCG 2019)*, volume 129 of Leibniz International Proceedings in Informatics (LIPIcs), pp. 58:1–58:15, 2019.

[AD17] Karen L. Adair and Angela E. Douglas. "Making a Microbiome: The Many Determinants of Host-Associated Microbial Community Composition." *Current Opinion in Microbiology*, **35**:23–29, 2017.

[AS11] Henry Adams and Jan Segert. "Simplicial Complex Filtration Demonstrations in Mathematica.", 2011. Available at `https://github.com/henryadams/files/tree/main` (accessed 28 May 2025).

[BB04] Lars Brabyn and Ross Barnett. "Population Need and Geographical Access to General Practitioners in Rural New Zealand." *New Zealand Medical Journal*, **117**(1199):1–13, 2004.

[BB06] Lars Brabyn and Paul Beere. "Population Access to Hospital Emergency Departments and the Impacts of Health Reform in New Zealand." *Health Informatics Journal*, **12**(3):227–237, 2006.

[BBG18] Hugo Barbosa, Marc Barthelemy, Gourab Ghoshal, Charlotte R. James, Maxime Lenormand, Thomas Louail, Ronaldo Menezes, José J. Ramasco, Filippo Simini, and Marcello Tomasini. "Human Mobility: Models and Applications." *Physics Reports*, **734**:1–74, 2018.

[BBH06] Raymond C. Browning, Emily A. Baker, Jessica A. Herron, and Rodger Kram. "Effects of Obesity and Sex on the Energetic Cost and Preferred Speed of Walking." *Journal of Applied Physiology*, **100**(2):390–398, 2006.

[Bor48] Karol Borsuk. "On the Imbedding of Systems of Compacta in Simplicial Complexes." *Fundamenta Mathematicae*, **35**(11):217–234, 1948.

[BPP20] Elisa C. Baek, Mason A. Porter, and Carolyn Parkinson. "Social Network Analysis for Social Neuroscientists." *Social Cognitive and Affective Neuroscience*, **16**(8):883–901, 2020.

[Bul24] Francesco Bullo. *Lectures on Network Systems*. Kindle Direct Publishing, 1.7th edition, 2024. Available at `https://fbullo.github.io/lns`.

[BYM22] Agnese Barbensi, Hee Rhang Yoon, Christian Degnbol Madsen, Deborah O. Ajayi, Michael P. H. Stumpf, and Heather A. Harrington. "Hypergraphs for Multiscale Cycles in Structured Data." *arXiv:2210.07545*, 2022.

[CHM22] Justin Curry, Haibin Hang, Washington Mio, Tom Needham, and Osman Berat Okutan. "Decorated Merge Trees for Persistent Homology." *Journal of Applied and Computational Topology*, **6**:371–428, 2022.

[CHP19] M. Keith Chen, Kareem Haggag, Devin G. Pope, and Ryne Rohla. "Racial Disparities in Voting Wait Times: Evidence from Smartphone Data." Working Paper 26487, National Bureau of Economic Research, 2019. Available at SSRN: `https://ssrn.com/abstract=3492890`.

[CHT22] Emil Dalgaard Christensen, Mathis Hjort Hjelmsø, Jonathan Thorsen, Shiraz Shah, Tamsin Redgwell, Christina Egeø Poulsen, Urvish Trivedi, Jakob Russel, Shashank Gupta, Bo L. Chawes, Klaus Bønnelykke, Søren Johannes Sørensen, Morten Arendt Rasmussen, Hans Bisgaard, and Jakob Stokholm. "The Developing Airway and Gut Microbiota in Early Life is influenced by Age of Older Siblings." *BMC Microbiome*, **10**(1):106, 2022.

[CIM24] Wenping Cui, Robert Marsland III, and Pankaj Mehta. "Les Houches Lectures on Community Ecology: From Niche Theory to Statistical Mechanics." *arXiv:2403.05497*, 2024.

[CJ23] Padraig Corcoran and Christopher B. Jones. "Topological Data Analysis for Geographical Information Science Using Persistent Homology." *International Journal of Geographical Information Science*, **37**(3):712–745, 2023.

[CSC19] Daniel F. R. Cleary, Thomas Swierts, Francisco J. R. C. Coelho, Ana R. M. Polónia, Yusheng M. Huang, Marina R. S. Ferreira, Sumaitt Putchakarn, Luis Carvalheiro, Esther van der Ent, Jinn-Pyng Ueng, Newton C. M. Gomes, and Nicole J. de Voogd. "The Sponge Microbiome Within the Greater Coral Reef Microbial Metacommunity." *Nature Communications*, **10**(1):1644, 2019.

[DW22] Tamal K. Dey and Yusu Wang. *Computational Topology for Data Analysis*. Cambridge University Press, Cambridge, UK, New edition, 2022.

[Ede05] Leah Edelstein-Keshet. *Mathematical Models in Biology*. Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 2005.

[EH10] Herbert Edelsbrunner and John Harer. *Computational Topology: An Introduction*. American Mathematical Society, Providence, RI, USA, 2010.

[Fel68] William Feller. *An Introduction to Probability Theory and Its Applications. Volume I*. John Wiley & Sons, Inc,, Oxford, UK, 3rd edition, 1968.

[Fen25] Michelle Feng. "Interpreting Topology in the Context of Social Science." In Heather Z. Brooks, Michelle Feng, Mason A. Porter, and Alexandria Volkening, editors, *Mathematical and Computational Methods for Complex Social Systems*, volume 80 of Proceedings of Symposia in Applied Mathematics. American Mathematical Society, Providence, RI, USA, 2025.

[FHP22] Michelle Feng, Abigail Hickok, and Mason A. Porter. "Topological Data Analysis of Spatial Systems." In Federico Battiston and Giovanni Petri, editors, *Higher-Order Systems*, Understanding Complex Systems, chapter 16, pp. 389–399. Springer International Publishing, Cham, Switzerland, 2022.

[Fow20] Stephan Fowler. "Why Do Nonwhite Georgia Voters Have to Wait in Line for Hours? Too Few Polling Places." NPR, 2020. Available at `https://www.npr.org/2020/10/17/924527679/why-do-nonwhite-georgia-voters-have-to-wait-in-line-for-hours-too-few-polling-pl`.

[FP20] Michelle Feng and Mason A. Porter. "Spatial Applications of Topological Data Analysis: Cities, Snowflakes, Random Structures, and Spiders Spinning Under the Influence." *Physical Review Research*, **2**:033426, 2020.

[FP21] Michelle Feng and Mason A. Porter. "Persistent Homology of Geospatial Data: A Case Study with Voting." *SIAM Review*, **63**(1):67–99, 2021.

[FZ24] Ori Friesen and Lori Ziegelmeier. "Understanding U.S. Racial Segregation Through Persistent Homology." *arXiv:2410.10886*, 2024.

[Goo] Google Developers. "Distance Matrix API." Available at `https://developers.google.com/maps/documentation/distance-matrix` (accessed 4–7 November 2021).

[GS03] James Gimpel and Jason Schuknecht. "Political Participation and Accessibility of The Ballot Box." *Political Geography*, **22**:471–488, 2003.

[Hat02] Allen Hatcher. *Algebraic Topology*. Cambridge University Press, Cambridge, UK, 2002.

[HJJ24] Abigail Hickok, Benjamin Jarman, Michael Johnson, Jiajie Luo, and Mason A. Porter. "Persistent Homology for Resource Coverage: A Case Study of Access to Polling Sites." *SIAM Review*, **66**(3):481–500, 2024.

[HK05] Moshe Haspel and H. Gibbs Knotts. "Location, Location, Location: Precinct Placement and the Costs of Voting." *The Journal of Politics*, **67**(2):560–573, 2005.

[HK12] Timothy L. Hawthorne and Mei-Po Kwan. "Using GIS and Perceived Distance to Understand the Unequal Geographies of Healthcare in Lower-Income Urban Neighbourhoods." *The Geographical Journal*, **178**(1):18–30, 2012.

[HLH05] Marcel Holyoak, Mathew A. Leibold, and Robert D. Holt. *Metacommunities: Spatial Dynamics and Ecological Communities*. University of Chicago Press, Chicago, IL, USA, 2005.

[HLS24] Drew Horton, Tom Logan, Daphne Skipper, and Emily Speakman. "Hundreds of Grocery Outlets Needed Across the United States to Achieve Walkable Cities." *arXiv:2404.01209*, 2024.

[HNP22] Abigail Hickok, Deanna Needell, and Mason A. Porter. "Analysis of Spatial and Spatiotemporal Anomalies Using Persistent Homology: Case Studies with COVID-19 Data." *SIAM Journal on Mathematics of Data Science*, **4**(3):1116–1144, 2022.

[HWC22] Kaijian Hou, Zhuo-Xun Wu, Xuan-Yu Chen, Jing-Quan Wang, Dongya Zhang, Chuanxing Xiao, Dan Zhu, Jagadish B Koya, Liuya Wei, Jilin Li, and Zhe-Sheng Chen. "Microbiota in Health and Diseases." *Nature Signal Transduction and Targeted Therapy*, **7**(1):135, 2022.

[Kan19] Henry Kanengiser. "In New York, Where You Live Can Determine How Hard It Is To Vote." *City Limits*, 2019. Available at `https://citylimits.org/2019/04/25/nyc-polling-place-shortage-inequality/`.

[KD24] Teeratorn Kadeethum and Christine Downs. "Harnessing Machine Learning and Data Fusion for Accurate Undocumented Well Identification in Satellite Images." *Remote Sensing*, **16**(12):2116, 2024.

[KSC22] Chong-Su Kim, Go-Eun Shin, Yunju Cheong, Ji-Hee Shin, Dong-Mi Shin, and Woo Young Chun. "Experiencing Social Exclusion Changes Gut Microbiota Composition." *Nature Translational Psychiatry*, **12**(1):254, 2022.

[KW24] Jakini Auset Kauba and Thomas Weighill. "Topological Analysis of U.S. City Demographics." *La Matematica*, **3**(4):1400–1425, 2024.

[LHM04] Mathew A. Leibold, Marcel Holyoak, Nicolas Mouquet, Priyanga Amarasekare, Jonathan M. Chase, Martha F. Hoopes, Robert D. Holt, Jonathan B. Shurin, Richard Law, David Tilman, Michel Loreau, and Andrew Gonzalez. "The Metacommunity Concept: A Framework for Multi-Scale Community Ecology." *Ecology Letters*, **7**(7):601–613, 2004.

[LMG03] Michel Loreau, Nicolas Mouquet, and Andrew Gonzalez. "Biodiversity as Spatial Insurance in Heterogeneous Landscapes." *Proceedings of the National Academy of Sciences of the United States of America*, **100**(22):12765–12770, 2003.

[LRR23] Brian A. Lerch, Akshata Rudrapatna, Nasser Rabi, Jonas Wickman, Thomas Koffel, and Christopher A. Klausmeier. "Connecting Local and Regional Scales with Stochastic Metacommunity Models: Competition, Ecological Drift, and Dispersal." *Ecological Monographs*, **93**(4):e1591, 2023.

[LTH21] Lu Li, Connor Thompson, Gregory Henselman-Petrusek, Chad Giusti, and Lori Ziegelmeier. "Minimal Cycle Representatives in Persistent Homology Using Linear Programming: An Empirical Study with User's Guide." *Frontiers in Artificial Intelligence*, **4**:681117, 2021.

[LWF03] Xiang-Yang Li, Peng-Jun Wan, and Ophir Frieder. "Coverage in Wireless Ad Hoc Sensor Networks." *IEEE Transactions on Computers*, **52**(6):753–763, 2003.

[Mac05] David MacKay. *Information Theory, Inference, and Learning Algorithms*. Cambridge Univesity Press, Cambridge, UK, 2005.

[MHJ08] Joël Mossong, Niel Hens, Mark Jit, Philippe Beutels, Kari Auranen, Rafael Mikolajczyk, Marco Massari, Stefania Salmaso, Gianpaolo Scalia Tomba, Jacco Wallinga, Janneke Heijne, Malgorzata Sadkowska-Todys, Magdalena Rosinska, and W. John Edmunds. "Social Contacts and Mixing Patterns Relevant to the Spread of Infectious Diseases." *PLoS Medicine*, **5**(3):e74, 2008.

[MKP01] S. Meguerdichian, F. Koushanfar, M. Potkonjak, and M. B. Srivastava. "Coverage Problems in Wireless Ad-Hoc Sensor Networks." In *Proceedings IEEE INFOCOM 2001. Conference on Computer Communications. Twentieth Annual Joint Conference of the IEEE Computer and Communications Society (Cat. No.01CH37213)*, volume 3, pp. 1380–1387, 2001.

[ML02] Nicolas Mouquet and Michel Loreau. "Coexistence in Metacommunities: The Regional Similarity Hypothesis." *The American Naturalist*, **159**(4):420–426, 2002.

[ML03] Nicolas Mouquet and Michel Loreau. "Community Patterns in Source-Sink Metacommunities." *The American Naturalist*, **162**(5):544–557, 2003.

[MSB18] Elizabeth Theresa Miller, Richard Svanbäck, and Brendan J.M. Bohannan. "Microbiomes as Metacommunities: Understanding Host-Associated Microbes through Metacommunity Ecology." *Trends in Ecology & Evolution*, **33**(12):926–935, 2018.

[New18] Mark Newman. *Networks*. Oxford University Press, Oxford, UK, second edition, 2018.

[OF03] Stanley J. Osher and Ronald Fedkiw. *Level Set Methods and Dynamic Implicit Surfaces*. Springer-Verlag, Heidelberg, Germany, 2003.

[Ope] OpenStreetMap contributors. "OpenStreetMap", 2021. Version 1.1.1 (accessed 19–22 January 2022).

[OPT17] Nina Otter, Mason A. Porter, Ulrike Tillmann, Peter Grindrod, and Heather A. Harrington. "A Roadmap for the Computation of Persistent Homology." *EPJ—Data Science*, **6**:17, 2017.

[OT24] Erin O'Neil and Sarah Tymochko. "Evaluating Cooling Center Coverage Using Persistent Homology of a Filtered Witness Complex." *arXiv:2410.09067*, 2024.

[PG16] Mason A. Porter and James P. Gleeson. *Dynamical Systems on Networks: A Tutorial*, volume 4. Springer International Publishing, Cham, Switzerland, 2016.

[PWB06] Jamie Pearce, Karen Witten, and Phil Bartie. "Neighbourhoods and Health: A GIS Approach to Measuring Community Resource Accessibility." *Journal of Epidemiology and Community Health*, **60**(5):389–395, 2006.

[RAT21] Aura Raulo, Bryony E. Allen, Tanya Troitsky, Arild Husby, Josh A. Firth, Tim Coulson, and Sarah C. L. Knowles. "Social Networks Strongly Predict the Gut Microbiota of Wild Mice." *The ISME Journal*, **15**(9):2601–2613, 2021.

[Sch21] Rachel Schnalzer. "Traffic is Terrible Again. Here's How to Get it Closer to Spring 2020 Levels." *Los Angeles Times*, 2021. Available at `https://www.latimes.com/business/story/2021-07-22/los-angeles-traffic-congestion-commute-pandemic`.

[SG06] Vin de Silva and Robert Ghrist. "Coordinate-Free Coverage in Sensor Networks with Controlled Boundaries Via Homology." *The International Journal of Robotics Research*, **25**(12):1205–1222, 2006.

[SG07] Vin de Silva and Robert Ghrist. "Coverage in Sensor Networks via Persistent Homology." *Algebraic and Geometric Topology*, **7**(1):339–358, 2007.

[SHJ20] Amar Sarkar, Siobhán Harty, Katerina V.-A. Johnson, Andrew H. Moeller, Elizabeth A. Archie, Laura D. Schell, Rachel N. Carmody, Timothy H. Clutton-Brock, Robin I. M. Dunbar, and Philip W. J. Burnet. "Microbial Transmission in Animal Social Networks and the Social Microbiome." *Nature Ecology & Evolution*, **4**(8):1020–1035, 2020.

[SHP16] Bernadette J. Stolz, Heather A. Harrington, and Mason A. Porter. "The Topological "Shape" of Brexit." *arXiv:1610.00752*, 2016.

[SMH24] Amar Sarkar, Cameron J. A. McInroy, Siobhán Harty, Aura Raulo, Neil G. O. Ibata, Mireia Valles-Colomer, Katerina V.-A. Johnson, Ilana L. Brito, Joseph Henrich, Elizabeth A. Archie, Luis B. Barreiro, Francesca S. Gazzaniga, B. Brett

Finlay, Eugene V. Koonin, Rachel N. Carmody, and Andrew H. Moeller. "Microbial Transmission in the Social Microbiome and Host Health and Disease." *Cell*, **187**(1):17–43, 2024.

[SSH19] Simon Scarr, Manas Sharma, and Marco Hernandez. "Roads, Boats and Elephants: How India Mobilised a Million Polling Stations." *Reuters*, 2019. Available at `https://graphics.reuters.com/INDIA-ELECTION-STATIONS/010092FY33Z/index.html`.

[SW] Christopher Stover and Eric W. Weisstein. "Einstein Summation." From MathWorld—A Wolfram Web Resource. Available at `https://mathworld.wolfram.com/EinsteinSummation.html` (accessed 23 April 2025).

[TBB15] Jenny Tung, Luis B Barreiro, Michael B Burns, Jean-Christophe Grenier, Josh Lynch, Laura E Grieneisen, Jeanne Altmann, Susan C Alberts, Ran Blekhman, and Elizabeth A Archie. "Social Networks Predict Gut Microbiome Composition in Wild Baboons." *eLife*, **4**:e05224, 2015.

[TGD20] Patrick L. Thompson, Laura M. Guzman, Luc De Meester, Zsófia Horváth, Robert Ptacnik, Bram Vanschoenwinkel, Duarte S. Viana, and Jonathan M. Chase. "A Process-Based Metacommunity Framework Linking Local and Regional Scale Community Ecology." *Ecology Letters*, **23**(9):1314–1329, 2020.

[The21] The White House. "FACT SHEET: President Biden Announces 90% of the Adult U.S. Population will be Eligible For Vaccination and 90% Will Have a Vaccination Site Within 5 Miles of Home by April 19." Available at `https://www.whitehouse.gov/briefing-room/statements-releases/2021/03/29/fact-sheet-president-biden-announces-90-of-the-adult-u-s-population-will-be-eligible-for-vaccination-and-90-will-have-a-vaccination-site-within-5-miles-of-home-by-april-19/`, 2021.

[TMP12] Amanda L. Traud, Peter J. Mucha, and Mason A. Porter. "Social Structure of Facebook Networks." *Physica A: Statistical Mechanics and its Applications*, **391**(16):4165–4180, 2012.

[US] U.S. Census Bureau. "American Community Survey, 2015–2019 Estimates." Available at `https://data.census.gov/` (accessed 7 November 2021).

[VWS18] Ana M. Valdes, Jens Walter, Eran Segal, and Tim D. Spector. "Role of the Gut Microbiota in Nutrition and Health." *BMJ*, **361**:k2179, 2018.

[WYL17] Baohong Wang, Mingfei Yao, Longxian Lv, Zongxin Ling, and Lanjuan Li. "The Human Microbiota in Health and Disease." *Nature Signal Transduction and Targeted Therapy*, **3**(1):71–82, 2017.

[ZH05] Honghai Zhang and Jennifer C. Hou. "Maintaining Sensing Coverage and Connectivity in Large Sensor Networks." *Ad Hoc & Sensor Wireless Networks*, **1**:89–124, 2005.