Title
Assessing Interpretability Through Physical Model Analysis

Author
Wilkinson, Ryan

Publication Date
2022

UNIVERSITY OF CALIFORNIA
Los Angeles

Assessing Interpretability Through Physical Model Analysis

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

by

Ryan Wilkinson

2022

ABSTRACT OF THE DISSERTATION

Assessing Interpretability Through Physical Model Analysis

by

Ryan Wilkinson
Doctor of Philosophy in Mathematics
University of California, Los Angeles, 2022
Professor Marcus L. Roper, Chair

Understanding complex data requires some form of model analysis, whether it be with machine learning, statistical, or physical models. Such analysis is useful for producing predictions, identifying dynamics from noise, and understanding the system in question. Complex or "black box" models can accurately predict dynamics of the world around us, and are useful tools when matching data is the primary goal. However, these models fail at increasing understanding of the system of study.

This thesis is dedicated to the analysis of simple models, and when, how, and why they work to represent data. Herein, we analyze the application of the simple SIR epidemic model to complex epidemiological dynamics present in the COVID-19 pandemic in the U.S. We show that, despite the model's simplicity and apparently violated assumptions, it still has a place in matching and predicting real data, and we can learn key intuition from this fact. We also identify key contraction arrangements in the network-like slime mold *Physarum polycephalum* using a low Reynolds number flow model. These contraction arrangements, or modes, are key to understanding how, exactly, the organism gets its wide array of different behaviors. Our analysis places these modes in a simple, physically-understandable context that will allow researchers to connect measurable physical features to real complex behavior.

The dissertation of Ryan Wilkinson is approved.

Deanna M. Needell

Alexander Hoffmann

Jun Yin

Marcus L. Roper, Committee Chair

University of California, Los Angeles

2022

To my parents, who always encouraged me to follow my own path in life

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

First, I must thank my wonderful advisor, Marcus Roper, for putting me on the path to this advanced degree. Without his encouragement and guidance, I would not even be in graduate school, much less writing this. Marcus played an advisory role in all aspects of the chapters of the following dissertation. I would also like to thank my committee members Jun Yin, Alex Hoffman, and Deanna Needell. Thanks to my collaborators Karen Alim and Matthew Koziol, who contributed to the work in Chapter 4. Karen advised on all *Physarum polycephalum*-specific topics, and Matthew provided crucial programming support. Thank you to Martha and Brenda for their speedy help with administrative affairs.

I would like to thank Jodi O'Keefe, Steve Trombetta, and Wayne Hild, who went above and beyond to foster my interest in mathematics during my early education. Thanks also to Joan Griffin and John Deaderick for their vital role in developing my writing style. Finally, a huge thanks to my fiancée Makayla, whose support has always—and will always—help me in all aspects of life.

Chapter 2 is based on: Ryan Wilkinson and Marcus Roper. "Modeling Insights from COVID-19 Incidence Data: Part I – Comparing COVID-19 Cases Between Different-Sized Populations." *arXiv preprint arXiv:2211.09010* (2022)

Chapter 3 is based on: Ryan Wilkinson and Marcus Roper. "Modeling Insights from COVID-19 Incidence Data: Part II – Why are compartment models so accurate?" *arXiv preprint arXiv:2012.13424* (2022).

Chapter 4 is based on: Ryan Wilkinson, Matthew Koziol, Karen Alim, and Marcus Roper. "Flow Modes Provide a Quantification of Physarum Network Peristalsis." *Available at SSRN 4267724.*

This work is funded by the System and Integrative Biology Training Grant (NIH T32-GM008185), and by a Research Grant from HFSP (Ref.-No: RGP0001/2021).

VITA

2017	B.S. in Applied Mathematics, UCLA, Los Angeles

PUBLICATIONS

Ryan Wilkinson and Marcus Roper. “Modeling Insights from COVID-19 Incidence Data: Part I – Comparing COVID-19 Cases Between Different-Sized Populations.” *arXiv preprint arXiv:2211.09010* (2022).

Ryan Wilkinson and Marcus Roper. “Modeling Insights from COVID-19 Incidence Data: Part II – Why are compartment models so accurate?” *arXiv preprint arXiv:2012.13424* (2022).

Ryan Wilkinson, Matthew Koziol, Karen Alim, and Marcus Roper. “Flow Modes Provide a Quantification of Physarum Network Peristalsis.” *Available at SSRN 4267724*.

CHAPTER 1

Introduction

The universe is filled with complicated dynamics that are important to understand. From the way our body transports blood to survive, to the plethora of factors contributing to the warming climate, to the spread of COVID-19: we seek ways of describing and understanding the world around us. With the recent exponential boom in computing power and memory, we are increasingly able to store and analyze these dynamics in their entirety despite the staggering complexity and size of data sets. Our limitation now rests upon the methods we use to analyze such data sets.

Machine learning, for example, is a powerful and ubiquitous tool that produces highly accurate predictions that can capture complex systems with high fidelity. For example, machine learning can be used to translate speech to text (often used in modern-day smart phones) [TPS18], track objects in a video to avoid collisions [DCO18], and extract smooth trajectories from noisy data [BKS20]. The results from these algorithms are so good that they can seem almost magical—and they may as well be. Machine learning is commonly thought of as a “black box” that takes in data and outputs a neat, usually highly accurate result. The alchemy happening inside the black box, however, is notoriously difficult to make sense of, and provides no understanding of the data along the way.

Classical statistical models for understanding data are easier to understand than models from machine learning, but are still difficult and sometimes impossible to disentangle. In [FKW22], for example, a classical method known as Principal Component Analysis (PCA) was performed on image sequences of the organism *Physarum polycephalum*, a network-like organism known for its large array of different behaviors [NG08, NYT00, RMM16, BMV12, DLB10, TTS10]. PCA was meant to identify "modes" of contraction of *P. polycephalum* in an effort to understand how these modes correlated to a variety of behaviors, and it did just that. However, despite the successful identification of contraction modes, questions remained as to *why* these were the modes chosen; the statistics yielded no insight to this end. Another high-profile example of a statistical model that had an impact on the everyday lives of U.S. citizens is that employed by the IHME during the COVID pandemic [IM20b, IM20a], which used a purely statistical kernel to provide COVID forecasts that directly impacted federal policy. While effective at making these predictions, the model, again, yielded no insights as to which sociodemographic/dynamic patterns were contributing to the forecast.

To get a clear understanding of the *why* of our data, models that assign parameter values to physical and interpretable quantities are required. Even these, however, can be rendered opaque by model complexity. The epidemiological dynamics present in the COVID pandemic, for example, can be well-described by physical models (as in [BFM20]) that take into account varied subcommunity interaction structures to predict disease spread. The fits of these models, however, yield a large array of parameter values that is hard to describe concretely, although they are more interpretable still than statistical models.

In this dissertation, we seek to analyze the derivation of models with few parameters and their application to complex data in an endeavor to form simple conclusions and intuitions about complicated situations. We study two systems in the following: the COVID-19 pandemic in the U.S., and the flow/contraction dynamics of the slime mold *P. polycephalum*. Herein, we develop a rationale for the application of the mean-field SIR model to complex COVID data, and a physically-constrained modal analysis of the contractions of *P. polycephalum* and their resultant cytoplasmic flow.

The rest of the dissertation is organized into 3 parts and a concluding chapter:

Chapter 2: We compare COVID-19 case-curves between different US states, by clustering case surges between March 2020 and March 2021 into groups with similar dynamics. We advance the hypothesis that each surge is driven by a subpopulation of COVID-19 contacting individuals, and make detecting the size of that population a step within our clustering algorithm. Clustering reveals that case trajectories in each state conform to one of a small number (4-6) of archetypal dynamics. Our results suggest that, while the spread of COVID-19 in different states is heterogeneous, there are underlying universalities in the spread of the disease that may yet be predictable by models with reduced mathematical complexity. These universalities also prove to be surprisingly robust to school closures, which we choose as a common, but high social cost, public health measure.

Chapter 3: We study the SIR-compartment model, which is among the simplest models that describe the spread of a disease through a population. The model makes the unrealistic assumption that the population through which the disease is spreading is well-mixed. Although real populations have heterogeneities in contacts not represented in the SIR model, it nevertheless fits real U.S. state data well at multiple points throughout the pandemic. In this chapter, we demonstrate mathematically how closely the continuous SIR model approximates a model which includes heterogeneous contacts, and provide insight onto how one can interpret parameters gleaned from regression in the context of heterogeneous dynamics.

Chapter 4: We study *Physarum polycephalum*, a foraging, network-forming organism known for its ability to make complex decisions and maintain memory of past stimuli without use of a complex nervous system. Self-organized peristaltic flows within the network transport nutrients throughout the organism and initiate locomotion and morphological changes. A key step in understanding *P. polycephalum*'s ability to change behavior is therefore forming descriptors of this peristaltic flow. In this chapter, we develop a dynamic network-based method for describing organism-wide patterns of tube contractions from videos of *P. polycephalum.* Our tool provides robust readouts of the diversity of global modes of tube contraction that could occur within a given network, based on its geometry and topology, and sensitively identifies when global peristaltic patterns emerge and dissipate.

CHAPTER 2

Modeling Insights from COVID-19 Incidence Data: Part I - Comparing COVID-19 Cases Between Different-Sized Populations

2.1 Introduction

The different approaches that different US states have taken to controlling or mitigating the spread of COVID-19 have created a test-bed for evaluating public health responses to future diseases. It is natural to ask, and many headline writers have already [Woo21, Bro21], whether specific states have done better or worse than others. California and Florida have been objects of frequent comparison, because of their similar mild winter weather and highly diverse populations, but stark policy differences on school re-openings, masking and indoor dining. However, to compare number of COVID cases, or number of deaths, between different populations, it is necessary to normalize by some measure of population size. Simply dividing by the total state populations, shows slightly larger case rates and much higher death rates in Florida (Table 2.1). By dividing by the total population of the state, we effectively treat the population of the state as a single entity, but both states have highly heterogeneous distributions of cases 2.1: in Florida both cases and case rates are concentrated in Miami-Dade county, while in California, case rates are highest in two low density rural counties (Imperial County and King's County), though the greatest number of cases occurs again in a single large metropolitan area (Los Angeles). Dividing the total number of cases (dominated by a few urban hotspots) by a total state population leads to a misleading picture of COVID-19 incidence. When comparing COVID incidences between different populations, we would like the population size that we normalize by to reflect a well-mixed subpopulation of individuals with similar levels of exposure and susceptibility to the disease. However, isolating these populations is not straightforward: even we study a single metropolitan area (e.g. Los Angeles county), we find widely different COVID case rates between different neighborhoods, separated only by miles (Fig 2.1).

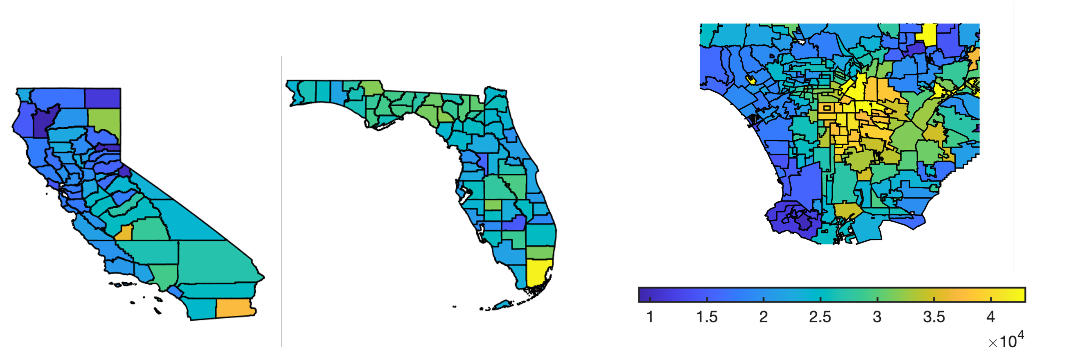


Figure 2.1: Normalization of cases by total population of state masks small spatial scale heterogeneities in case rates. In California, Kings and Imperial Counties have highest case rates, but numerically largest number of cases is in Los Angeles County (left). Miami-Dade county has highest case numbers and rates in FL (middle). Case rates in Downtown and East Los Angeles County neighborhoods are 2-3 times higher than in West and South Bay neighborhoods (right). Shown: cumulative data number of COVID cases per 100,000 individuals, on April 24, 2022. Sources: FL and CA county data: New York Times, COVID-19 dashboard, LA neigborhood data, Los Angeles Times, COVID-19 dashboard.

Among mathematical models that have been deployed to predict COVID spread, and to assist with the allocation of resources, some, such as agent based and network models, specifically address the role of population heterogeneities in shaping the spread of the disease [KE05, BBT20, MPV02, ACG20, BGM07, BRP10, Kee05, Li20, VM07]. Other models upscale heterogeneous populations into single, well-mixed, groups of individuals, trading off the flexibility in forming the dynamics of interaction that a more complex model affords,

	California	Florida
Population	39.19 M	21.48 M
Cases	9,237,030	5,963,941
Deaths	90,706	74,056
Case rate	23,570	27,380
Death rate	231	340

Table 2.1: Cumulative numbers of COVID-19 cases and deaths in California and Florida, on April 22nd, 2022. Naive comparisons between states are based on case and death rates /100,000 individuals, calculated by dividing by the total state population. Source: New York Times COVID-19 dashboard.

against a smaller set of parameters are easier to interpret and to fit against data [BFM20, SSV13, TL20, Het00a]. A third, influential, class of models is purely data fitted, with no mechanistic interpretation about how individuals are interacting or transmitting the disease [IM20b, IM20a].

Here we seek to shed light on both problems: 1. how to compare COVID cases among two populations of two sizes, and 2. what is the appropriate level of complexity to use in a mathematical model describing the spread of COVID-19 within a population. Specifically, we seek to compare the COVID-19 incidence curves between different US states, as paradigms of large heterogeneous, incompletely mixed populations. Comparing the data reveals that there are natural normalizations for case curves, which we interpret as the size of the subpopulation that has contact with COVID-infected individuals. The second goal of our analysis is to determine how diverse COVID-19 incidence time courses truly are: whether each State follows a completely distinct disease trajectory, or whether there are common families of time courses representing quantitatively similar dynamics of disease spread that could put an upper bound on the complexity of models needed to predict the course of the epidemic. Our third goal is to make trend-conscious comparisons that could be used to quantitatively compare the effectiveness of public health control measures practiced in different populations based on the time course of the pandemic rather than on single point in time measurements, such as those given in Table 2.1.

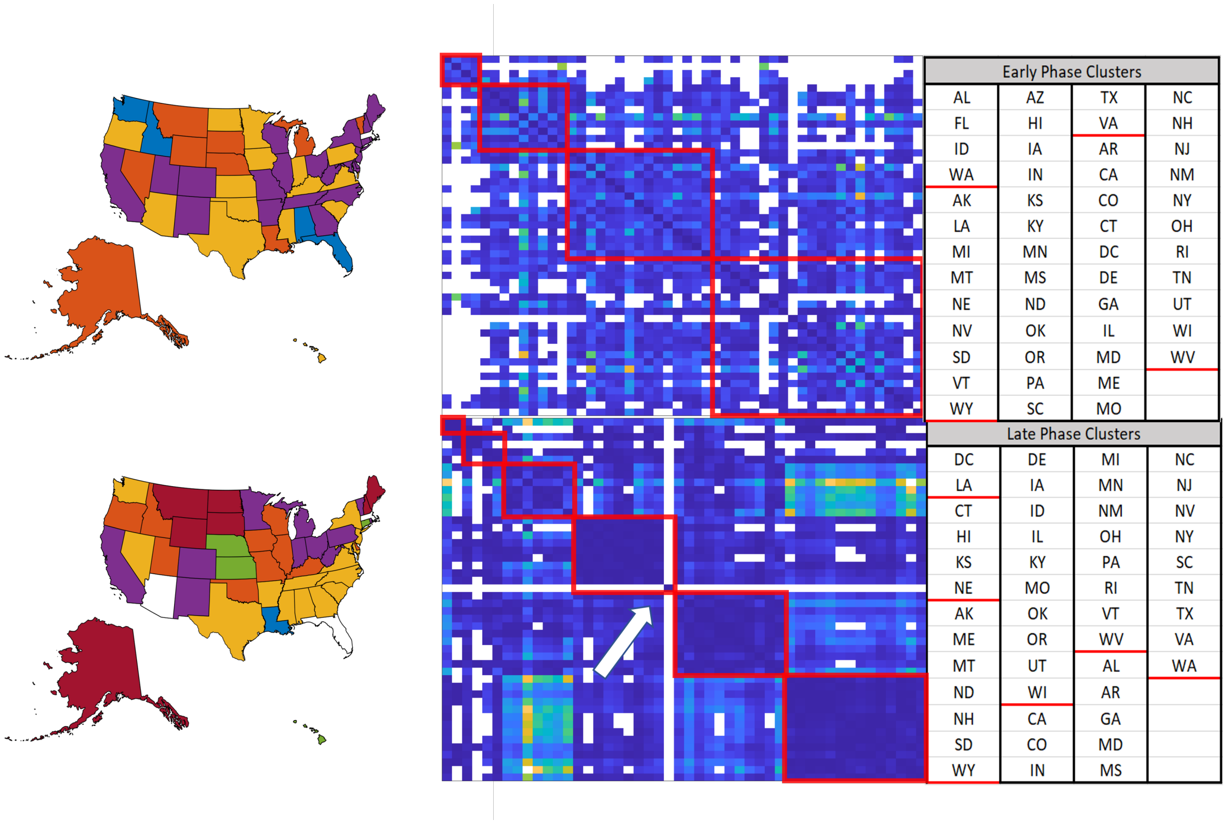
To this end, we analyzed one year of cumulative U.S. State case data (from March 2020 to March 2021) taken from The COVID Tracking Project[The], and isolated the earliest and last complete waves of spread contained in these data. We clustered case curves of similar shape, normalizing each case curve by a population-size that was allowed to different from the overall state population.These clusters revealed surprising case curve homologies that grouped states with very different voting preferences, urban/rural densities, and demographies. Comparisons within clusters gave estimates of the sizes of the populations among which active COVID transmission is occurring, whereas comparisons between clusters may allow for the evaluation of the effectiveness of public health measures controlling spread.

2.2 Results

2.2.1 U.S. State Clustering

Using methods outlined in Appendix A, we first used the WPGMA algorithm [Sok58] to select 14 groups of case curves from the entire corpus of data with early and late phases included. The clustering immediately separated early phase from late phase with no exceptions. 6 such clusters included early phase case curves, and 8 included late phase. Plotting the clusters, we noticed that 3 early phase clusters could be readily combined into a single super cluster. Moreover, we noticed that Wisconsin (Fig 2.2 white arrow) could be combined into another cluster, and that two clusters initially of two states (Connecticut and Hawaii, Kansas and Nebraska) may be aggregated into a single cluster of four states. This curation reduced the number of clusters to 4 early phase clusters and 6 late phase clusters.

Fig 2.3 suggests that all states in their initial case surge exhibit very similar behavior on the log scale, with minor variations towards the end of the surge as more complicated



Early Phase Clusters			
AL	AZ	TX	NC
FL	HI	VA	NH
ID	IA	AR	NJ
WA	IN	CA	NM
AK	KS	CO	NY
LA	KY	CT	OH
MI	MN	DC	RI
MT	MS	DE	TN
NE	ND	GA	UT
NV	OK	IL	WI
SD	OR	MD	WV
VT	PA	ME	
WY	SC	MO	

Late Phase Clusters			
DC	DE	MI	NC
LA	IA	MN	NJ
CT	ID	NM	NV
HI	IL	OH	NY
KS	KY	PA	SC
NE	MO	RI	TN
AK	OK	VT	TX
ME	OR	WV	VA
MT	UT	AL	WA
ND	WI	AR	
NH	CA	GA	
SD	CO	MD	
WY	IN	MS	

Figure 2.2: U.S. states are clustered in their early and late case surges via our distance metric. The dissimilarity heatmap (center top/bottom) represents the raw dissimilarity data, with white pixels indicating incompatible State curves due to incommensurate surge lengths. Red squares outline the clusters that were decided with a combination of hierarchical WPGMA clustering [Sok58] and manual curation, and correspond to the listed clusters in the chart (right top/bottom). The super cluster created from manual curation in the early phase is the largest cluster, and is the bottom-right-most square in the heatmap. Massachusetts is excluded from analysis in both early and late phases due to poor data quality, and Arizona and Florida were identified as out groups for late phase clustering.

dynamics emerge. Each curve begins as a straight line and appears to be approaching level on the log scale, suggesting decelerating infection rates from an initial exponential surge in cases.

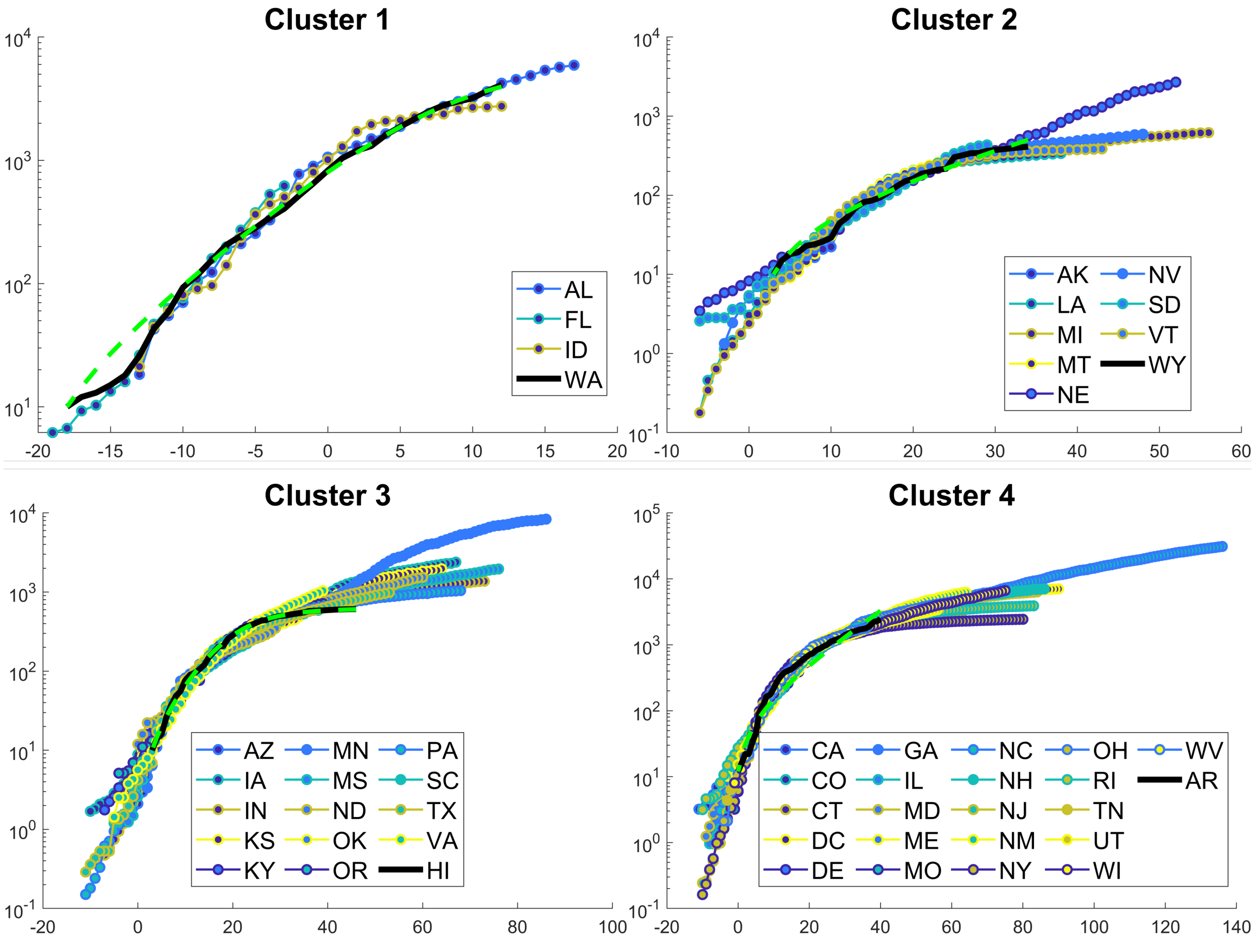


Figure 2.3: Clustered early phase cumulative case curves shown on log scale. A suitable reference curve is chosen based on average distance to other curves in its cluster, and other data is scaled and shifted to match the reference curve as closely as possible. The reference curve is named at the top of each panel and is shown as the solid black line in the plot. The SIR model fit to the reference curve is shown as a green dashed line. The horizontal axis of each plot represents time before or after the start of the reference curve's tenth case once the data have been shifted.

2.2.2 Detecting Effective COVID-contracting Population Size

When comparing case curves metrics between states, one usually scales the number of cases by total population size in order to measure per capita statistics rather than simple raw

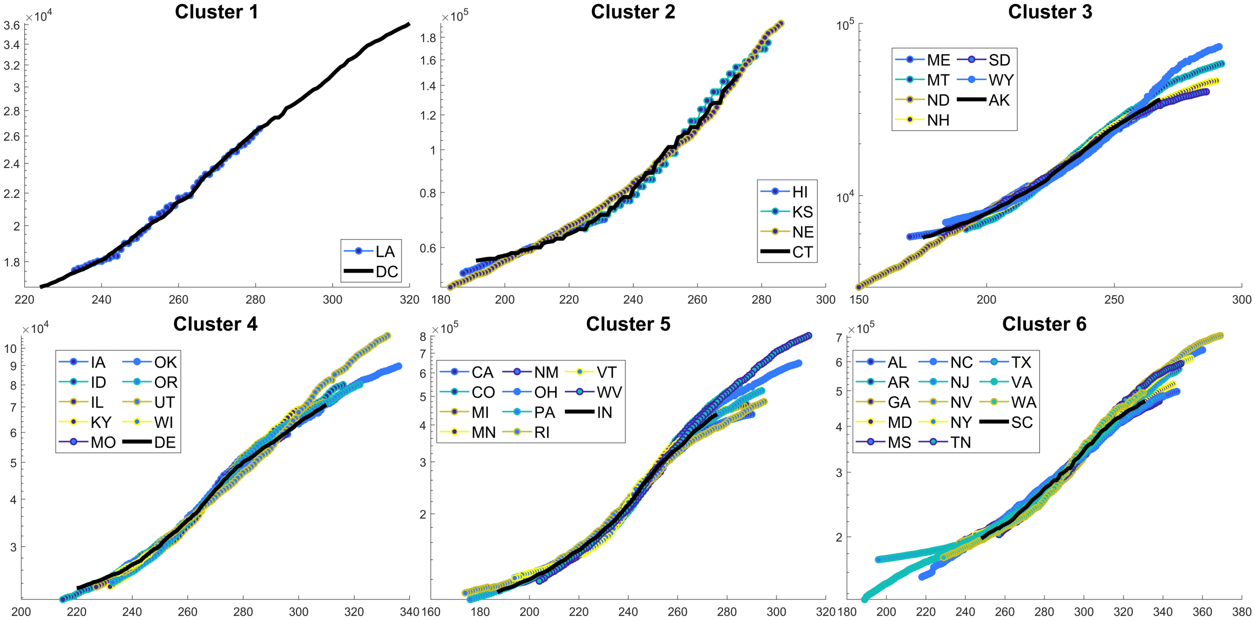


Figure 2.4: Clustered late phase cumulative case curves are shown on the log scale. A suitable reference curve is chosen based on average distance to other curves in its cluster, and other data is scaled and shifted to match the reference curve as closely as possible. The reference curve is named at the top of each panel and is shown as the solid black line in the plot. The horizontal axis of each plot represents time before or after the start of the reference curve's tenth case once the data have been shifted.

data. However, on the relatively large U.S. State scale, normalizing by population may not be entirely appropriate, because doing so implicitly assumes that the entire population contributes to the growth in cases, and ignores the complex reality wherein cases may be localized in hot spots whose size may not correlate closely to the population as a whole. When analyzing such statistics, then, it may therefore be helpful to focus on the cities, counties, or even neighborhoods in which mixing of individuals is causing COVID transmission. By allowing population sizes to be freely rescaled when comparing states, our alignments can be used to estimate the relative sizes of COVID-impacted populations between any two given states, as mentioned in section SII. We tested whether the ratios of population size

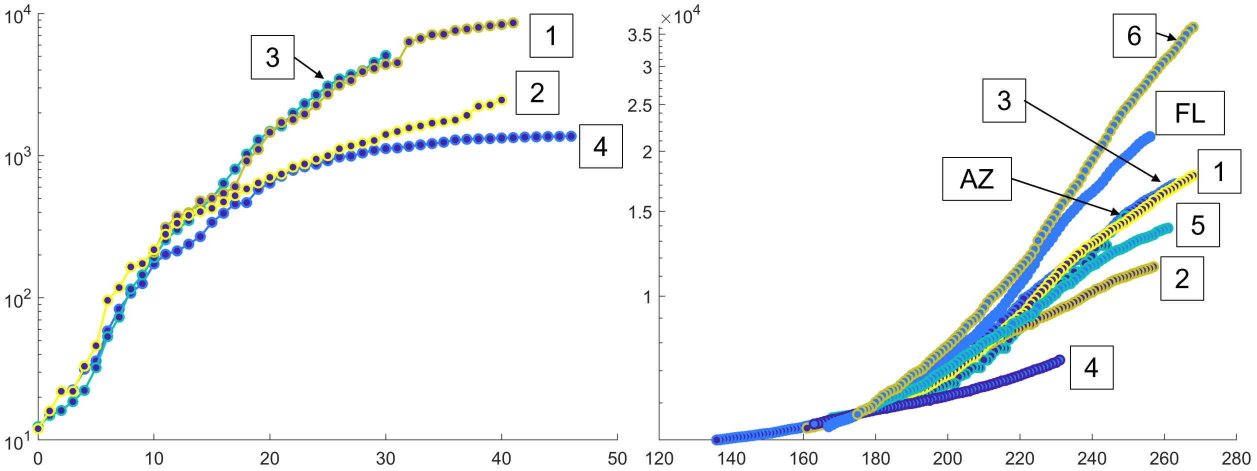


Figure 2.5: Representative curves from each respective cluster are compared with alignment biased toward the beginning of the curve in order to emphasize disparate evolution of dynamics, shown on log scale. Number labels correspond to cluster IDs, and match those shown in Figures 2.3 and 2.4. Left: Early phase case curves. Right: Late phase case curves, including Florida and Arizona, which are excluded from previous analyses.

of states also predict the ratios of their COVID-contracting populations (Table 2.2), finding that there is a weak R^2 of 0.31. A similar score was seen when total State population was replaced by urban population (given by the total population of the largest 3 cities in each State (Table 2.2, $R^2 = 0.2005$)), reflecting, perhaps the dominant role played in many states by transmission of COVID within urban populations. However, our analysis also exposed some notable exceptions. For example, our clustering ratio between Florida and the other states in its cluster were routinely lower than the actual population ratios of the states in that cluster, indicating that states matching Florida's case dynamics had to be scaled up less than the expected population ratio in order for the curves to align, suggesting that the COVID-contracting population was a smaller fraction of Florida's total population than for states with similar case dynamics. Conversely, Idaho had the opposite result (higher clustering ratio versus population ratio), indicating that it had a larger COVID-contracting population fraction than states in the same cluster.

Correlation Tested	R^2 value
Clustering population ratio (log scale) vs. actual population ratio (log scale)	0.3169
Clustering population ratio (log scale) vs. population ratio of sum of 3 largest cities (log scale)	0.2005

Table 2.2: A summary of the basic linear regressions done on the data.

2.2.3 Searching for Variables to Explain Clustering

The clustering of states allows us to test hypotheses about which underlying variables, including both geographical proximity and similarities or divergences between public health measures adopted in different states. A lack of testing capacity—especially at the beginning of the COVID-19 pandemic—along with the high frequency of asymptomatic infections cloud estimates of the true prevalence of the disease. PCR test results of individuals on a cruise ship amid a COVID outbreak revealed high variability in test positivity rates between different testing methods [OYS21]. Antibody tests also likely depend on disease severity [HJP21], and as a result can prove to be ineffective at retrospective analysis of COVID cases. Even so, studies that have been done using antibody tests have revealed false negatives in all methods of PCR testing [KNS21]. Additionally, limited availability of COVID tests, particularly early in the pandemic, may mean that many individuals infected with COVID were not tested [LKK21]. The number of positive cases is therefore likely under-counted (although estimates vary for the extent to which cases have been under-counted [IR21, ND21, LKK21]), creating concern that the clusters identified in this study may be distorted or even dominated by different levels of testing coverage and different kinds of testing methods. We tested whether differences in testing coverage explained the different time dependencies of cases. We used percentage of positive tests as a proxy for coverage since this indicates when testing coverage is low: for example the test positivity rate reached over 40% in New York State during its first surge. We found that the average positivity rates within clusters were no more similar than would be expected under random grouping ($p = 0.5365$, by permutation test). We may therefore conclude that our clustering is not simply capturing differences in testing coverage.

We also performed our modified Mantel test on the data to see whether dissimilarity scores between states are correlated with physical distances between the population centroids of states being compared. Although there are few known examples of COVID hotspots spanning State boundaries, such similarity would be expected if two states are linked by high rates of migration. Some pairs of neighboring states showed expected close similarity: for example much of the Southeast, including Georgia, Alabama, and Tennessee (Fig 2.2), as well as Idaho, North, and South Dakota. Other clusters were made up of geographically distant states: for example, Connecticut, Hawaii, Nebraska, and Kansas all emerged within a single cluster (Fig 2.2). Accordingly, we checked whether close states tended to cluster together. The correlation between closeness of our clustering was low only $\rho = 0.0991$, but it is statistically significant ($p = 2.5486 \times 10^{-8}$, by Mantel test, Table 2.3). To further probe the issue, we also measured the average percentage of states in each cluster that share a bordering neighbor also in the cluster. We found that approximately 63% of states on average had a neighbor in their cluster. This number was tested it against 50,000 permutations of our clustering for a significance of $p = 1.0868 \times 10^{-5}$. Both tests indicate that our clustering did indeed tend to cluster close states more than a random clustering would, but also that closeness is not necessarily a very important factor deciding whether dynamics are similar.

Public health responses within states are strongly influenced by political climate and divisions between the responses of the two main political parties to a joint public health and economic crisis. We considered two measures of political climate: the party affiliation of the State governor at the beginning of the pandemic (excluding Washington D.C., which does not have a governor), and the percentage of voters who voted for the Republican candidate, Donald Trump, in the 2020 presidential elections. Neither measure was significantly correlated with State clustering (governor party, $p = 0.3068$, and Trump voting percentage, $p = 0.1453$).

Metric Tested	Test Statistic	p value
Population centroid distance	Pearson correlation	2.5486×10^{-8}
Absolute Trump voting percentage difference	Pearson correlation	0.1453
Majority party governorship status	Average percentage of most represented party in cluster	0.3068
States clustered together both in early and late phase	Average percentage of states in each cluster that are clustered together in both early and late phase	0.3893
States clustered with their bordering neighbors	Average percentage of states in each cluster that shared a neighbor in the cluster	1.0868×10^{-5}
Testing positivity rates	Average variance of testing positivity in each cluster	0.5365
School reopening dates	Average variance of school reopening dates	0.2516

Table 2.3: A summary of results from permutation tests (including our modified Mantel tests) done on our clustering. The left column describes a metric used to probe our clustering behavior. The p value in the right column refers to the percentage of time that the metric tested on a permuted version of our clustering exceeded that of our actual clustering. For each test, the number of permutations tested is 50,000.

Finally, we use school closures as a direct index of the level of social distancing enforced in each State. Some states (e.g. Florida and Rhode Island) adopted a single policy on keeping schools opened or closed; in other states, individual school districts or counties determined whether or not to offer in-person instruction. Hence, we used the largest school district in each State as a proxy for the entire State's response to the pandemic. We compiled data on when the largest school district in each State first opened some form of in-person learning, and found that whether or not schools were open during the last recorded surge did not correlate with our State clustering ($p = 0.2516$).

CHAPTER 3

Modeling Insights from COVID-19 Incidence Data: Part II - Why are compartment models so accurate?

3.1 Introduction

Differences in rates of contact and of susceptibility between individuals strongly affect both their likelihood of catching COVID-19 and their health outcomes once infected. Incorporating these heterogeneities into models of disease spread is essential to understand the differential impacts of the disease upon different subpopulations, such as Black or Brown Americans [Rey20, MMZ20], nursing home residents [YLI20], incarcerated [SPW20] and unvaccinated individuals [MVZ21]. As well as ensuring a more equitable understanding of the disease, baseline questions about whether, e.g. it was safe to reopen schools in the middle of the pandemic, can not be definitively answered without considering the different levels of vulnerability of the communities affected: students, their caretakers, teachers and school staff.

However, heterogeneity-capturing models contain many unknown parameters that are difficult to fit to real data, and are hard to interpret once fit. Accordingly, public health departments continue to make predictions about the progress of the epidemic and about the effectiveness of social distancing based on so-called well-mixed models (reviewed in [BFM20]), and these models can be made to fit existing case data very well. They suffer, however, from having assumptions that are too simplistic to honestly reflect complex social behaviors inherent in disease spread, and as such, parameter fits should be interpreted with caution. In the well-known SIR model [DH00], which is the focus of this paper, a group of susceptible individuals (S) transition via contact with infectious individuals to the infectious (I) group, and after some time recover or are removed (R). The compartments evolve according to:

$$\frac{\mathrm{d}S}{\mathrm{d}t} = -\beta \frac{SI}{N_{tot}} \ , \quad \frac{\mathrm{d}I}{\mathrm{d}t} = \beta \frac{SI}{N_{tot}} - \gamma I \ , \tag{3.1}$$

while $\frac{dR}{dt} = \gamma I$ ensures that, neglecting disease and other mortality, the total number of individuals $S + I + R = N_{tot}$ remains constant. The *susceptibility* coefficient β represents the number of infections caused by a single infected individual in an otherwise susceptible population in unit time. γ is a basic recovery rate. For the above equations to be valid the population should be well-mixed: everyone in the population interacts with everyone else at all times. Given geographic considerations and changes in mixing behavior during a pandemic, this assumption cannot be true. Accordingly, much modern epidemiological modeling, including of COVID-19 [BFM20, ACG20, TL20], has focused on the role of heterogeneous contacts, either by ramifying compartments or by using networks to model connections between individuals [MPV02, BBT20, Li20, KE05, EK02, Kee05, MPV02, BRP10, VM07, Yan08]. Although inclusion of heterogeneities can drastically affect e.g. thresholds for herd immunity [BBT20], the multiplication of parameters that occurs when heterogeneities are added to models makes it hard to validate these predictions against real data. Previous work has shown that under certain conditions heterogeneous models can be approximated by well-mixed models [KRM97, SSV13, BGM07], but there is limited data showing these conditions are met by real epidemics.

In chapter 2, we showed that the COVID-case curves from different US states and during different surges can be clustered into between 4 and 6 groups. The collapse of case curves from different states to a small number of master curves suggests, that in spite of heterogeneous COVID transmission rates and impacts, relatively low model complexity is needed to reproduce the overall growth and the decay in number of infectious cases in a surge. Alignment of case data from different states involved translating data in time, and rescaling number of cases by a population size that was detected during data clustering (chapter 2), and that we interpreted to be the size of the subpopulation through which COVID-transmission was occurring. It follows that any mathematical model to describe this data collapse, must have the same symmetries; and these symmetries are present within the SIR model [BFM20].

We therefore postulated that the SIR model, described above, might be able to describe some of the families of case dynamics. We fit the SIR model to data from the first COVID surge, specifically the cumulative number of infections detected in each US states from the beginning of their respective outbreaks in February or March of 2020, up to May 20, 2020 (by which time stay at home orders had been relaxed, changing the transmission rate of the disease [MHS20]). In total, the SIR model fits the US state data for 2 of the 4 initial COVID phase clusters, accounting for 26 of the 52 states and territories in the data set (see Fig 3.1 and Fig 3.2) of this paper for examples of early COVID SIR fits). The SIR model performed much better for the most recent surge in the Omicron variant with few exceptions, closely conforming to the true case trajectory. Representative fits from SIR models to the Omicron variant are shown in Fig 3.3. In this paper, we examine two questions: First, given its neglect of heterogeneities in population contacts, why does the well-mixed SIR differential equation fit real pandemic data so well? Second, when we fit the SIR model to real observational data, can we interpret its fitted parameters, when the interaction processes that are parameterized are themselves not realistic? We find that the fitted β parameter implicitly accounts for heterogeneous interaction dynamics, and the N_{tot} parameter reveals an effective epidemic population size. We then apply our findings to the surprisingly well-fit omicron variant surge.

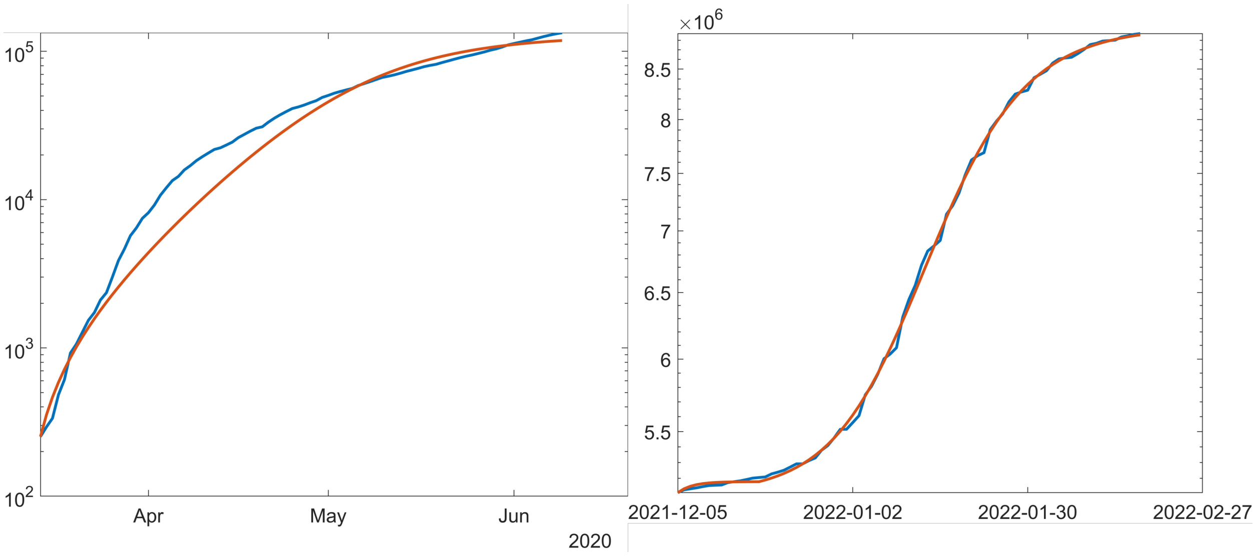


Figure 3.1: California's cumulative COVID case data (blue) is reasonably approximated by the SIR model (orange) both in its initial stage and during the Omicron surge (left and right panels, respectively). Initial COVID data is highly dependent on testing capacity which was highly variable during the initial wave of COVID, so goodness of fit of the initial wave is best gauged by the fit to the latter part of the data.

3.2 Results

3.2.1 A Heterogeneous Extension

To examine why the SIR model fits real pandemic data at all, we consider a multi-population SIR model that allows for populations to be heterogeneous in their contacts and susceptibilities. We then interrogate simulations resulting from this model to examine how important such heterogeneity is in the dynamics, and what fitting such a model with a homogeneous model means mathematically.

The model assumes that the population is split into K subpopulations, with population sizes N_i, $i = 1, 2, \ldots, K$, that interact with each other with contact rate β_{ij}, i.e. β_{ij} is the number of infections a single infected individual in subpopulation i could cause in subpopu-

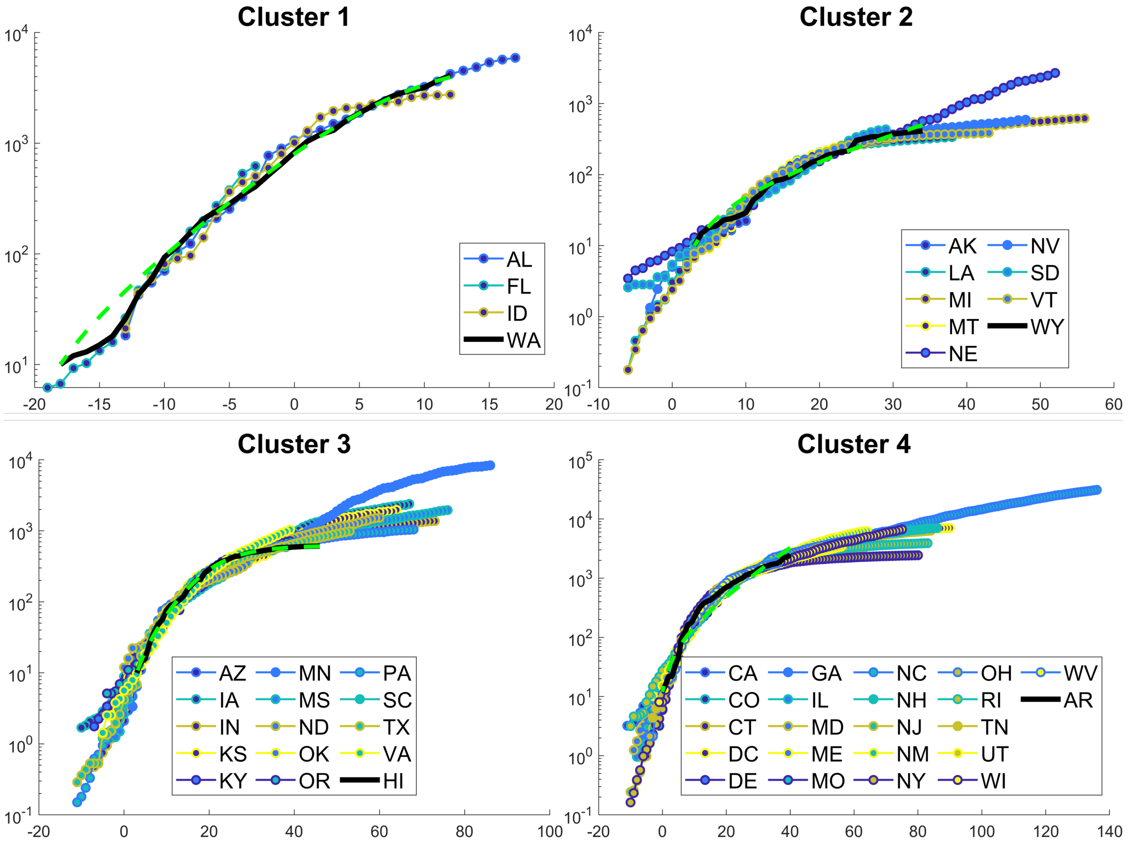


Figure 3.2: Shown are the computed clusters grouped together by the methods described in chapter 2. The least-fitting cluster (in the authors' opinion) lining up with Arkansas still fits reasonably well when the fit at later stages is considered, especially when one keeps in mind that case data is highly unreliable and sensitive to stochastic effects at the beginning of the pandemic.

lation j, if j contains only susceptible individuals. Our subpopulations represent geographic or demographic partitions of a population of size N_{tot}. In this first treatment we neglect differences in recovery rate based on individual characteristics. The susceptible, infectious, and removed quantities of subpopulation i is given by S_i, I_i, and R_i respectively and also

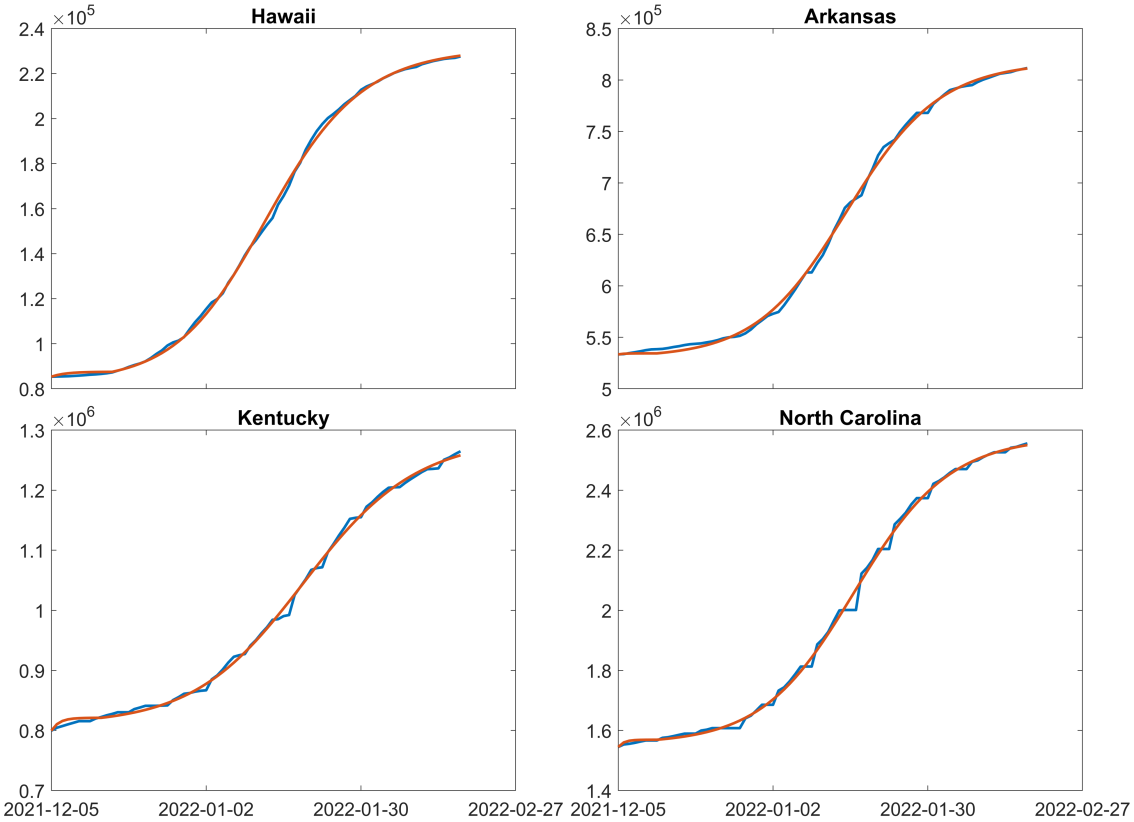


Figure 3.3: The SIR model (orange) fits Omicron surge data (blue) robustly for the vast majority of U.S. states. Shown are four examples of good fits for states with wildly different social/demographic structures and differing smoothness of data reporting.

write $\sum_i S_i = S$, $\sum_i I_i = I$, $\sum_i R_i = R$, and $\sum_i N_i = N_{tot}$. Consider the model:

$$\begin{aligned}
\frac{\mathrm{d}S}{\mathrm{d}t} &= -\frac{\tilde{\beta}}{N_{tot}} SI - \sum_{i=1}^{K}\sum_{j=1}^{K}\left(\frac{\beta_{ij}}{N_j} - \frac{\tilde{\beta}}{N_{tot}}\right) S_i I_j \\
\frac{\mathrm{d}I}{\mathrm{d}t} &= \frac{\tilde{\beta}}{N_{tot}} SI - \gamma I + \sum_{i=1}^{K}\sum_{j=1}^{K}\left(\frac{\beta_{ij}}{N_j} - \frac{\tilde{\beta}}{N_{tot}}\right) S_i I_j
\end{aligned} \tag{3.2}$$

To emphasize the similarity between the multi-population and the single-population SIR model we write the equation in a form with an SIR part, with arbitrary coefficient $\tilde{\beta}$ plus a residual. We obtain a type of best SIR fit by minimizing the L^2 norm of the ratio of the heterogeneous residual to the homogeneous term on an arbitrary interval (t_1, t_2), i.e. we minimize the size of the function

$$F(\tilde{\beta}, t) \equiv \frac{\sum_{i=1}^{K} \sum_{j=1}^{K} \left(\frac{\beta_{ij}}{N_j} - \frac{\tilde{\beta}}{N_{tot}}\right) S_i I_j}{\frac{\tilde{\beta}}{N_{tot}} SI} \tag{3.3}$$

This yields the unique minimum

$$\tilde{\beta} = N_{tot} \frac{\int_{t_1}^{t_2} \frac{\left(\sum_{i=1}^{K} \sum_{j=1}^{K} \frac{\beta_{ij}}{N_j} S_i I_j\right)^2}{\left(SI\right)^2} \, dt}{\int_{t_1}^{t_2} \frac{\sum_{i=1}^{K} \sum_{j=1}^{K} \frac{\beta_{ij}}{N_j} S_i I_j}{SI} \, dt} \tag{3.4}$$

We explore this quantity for a certain class of subpopulation interaction structure in Section (3.2.2) and (3.2.3).

3.2.2 Intermediate Mixed/Unmixed Models

We first consider the case where subpopulations either do not interact or else interact at identical rates:

$$\beta_{ij} = \beta\Big(\big(b\frac{N_j}{N_{tot}} + (1-b)\big)\delta_{ij} + b\frac{N_j}{N_{tot}}(1-\delta_{ij})A_{ij}\Big) \tag{3.5}$$

Here δ_{ij} is the Kronecker delta, b a mixing parameter ranging from 0 to 1, and A_{ij} is an adjacency matrix describing which subpopulations interact. Assume initially $A_{ij} \equiv 1$. When b is 0 the populations do not mix, and infections spread within but not between subpopulations, and when $b = 1$ the subpopulations mix completely, effectively merging into a single homogeneous population by construction of Eq. (3.5). Scaling β_{ij} by N_j ensures that interactions between subpopulations are proportionate to their sizes.

A homogeneous SIR model fits this model for $b \gtrsim 0.2$ (Fig (3.4)). Moreover, the $\tilde{\beta}$ gathered by Eq. (3.4) is very close to the β achieved by least squares model fitting to the full simulation, although unsurprisingly least squares fitting performs better at fitting the actual simulated curve for lower values of b. The agreement between $\tilde{\beta}$ and the fitted β is encouraging: for the consideration of real data, one can only deduce model parameters via some sort of fitting algorithm without knowledge of subcompartmental dynamics, but the result suggests that least squares fitting optimally estimates the susceptibility parameter to maximize the SIR part of the model relative to the interpopulation dynamics.

3.2.3 A Model with Network Structure

To investigate possible effects of subcommunity interaction structure on the homogeneous-like dynamics, we now model subpopulations whose interactions are prescribed by a network with adjacency matrix A_{ij}. $A_{ij} = 1$ indicates two subpopulations that maintain frequent contact with one another, such as a pair of communities which go to the same grocery store or school. For the purpose of analysis, we modeled random connections between subpopulations as random Erdős-Réyni networks parameterized by mean degree. The SIR model approximates the graph dynamics model above $b = 0.2$, with the fit improving as the mean degree of each node is increased (Fig (3.5)). Just as for a complete graph, $\tilde{\beta}$ agrees with β from least squares fits. Narrowing of error bars as mean degree of the random network increases express the decreasing importance of network structure as the network becomes more densely connected.

However, $\tilde{\beta}$ does not asymptote to $\beta = 0.2$ whether we use Eq. (3.4) or least squares fitting. Instead, increasing mixing allows the model to gain awareness of how sparse the connections between subpopulations are, ultimately causing $\tilde{\beta}$ to decrease. Fig (3.5) suggests that as the mean degree increases, we expect that rate of decrease to slow, and that $\tilde{\beta}$ will eventually asymptote to 0.2 as the network becomes complete (Fig (3.4)). Indeed, when all subpopulations have the same size, $\tilde{\beta}$ may be written as

$$\tilde{\beta} = \beta \frac{\int_{t_1}^{t_2} \left(\frac{K-(K-1)b}{K} \left(1 + K \sum_{i=1}^{K} \frac{S_i}{S} \left(\frac{I_i}{I} - \frac{1}{K}\right)\right) + b \frac{\sum_{i=1}^{K} \sum_{j=1}^{K} A_{ij} S_i I_j}{SI} \right)^2 dt}{\int_{t_1}^{t_2} \frac{K-(K-1)b}{K} \left(1 + K \sum_{i=1}^{K} \frac{S_i}{S} \left(\frac{I_i}{I} - \frac{1}{K}\right)\right) + b \frac{\sum_{i=1}^{K} \sum_{j=1}^{K} A_{ij} S_i I_j}{SI} \, dt} \tag{3.6}$$

Both numerator and denominator include a term which relies on graph structure and gains weight with the mixing parameter b and the covariance between the fraction of susceptibles and infecteds across subpopulations. The $\sum_{i=1}^{K} \frac{S_i}{S} \left(\frac{I_i}{I} - \frac{1}{K}\right)$ term represents a covariance

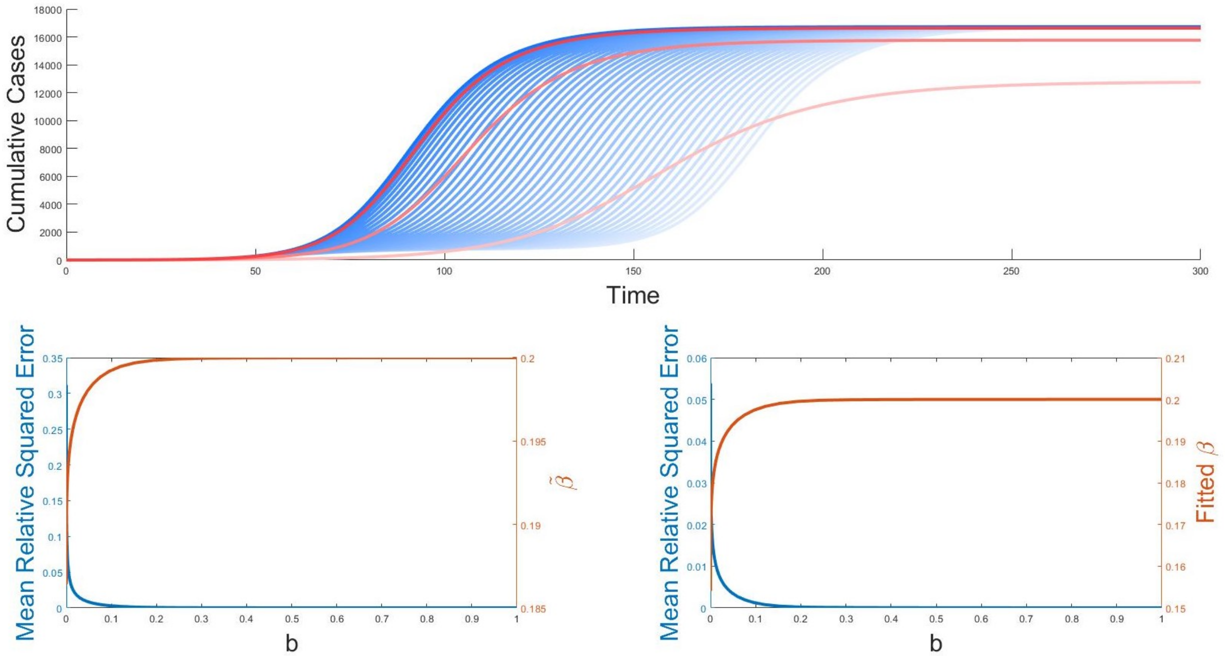


Figure 3.4: A single well-mixed population model can represent aggregate spread of disease through a linked set of subpopulations, even with relatively weak inter-subpopulation mixing. Top: Cumulative cases from the model in Eq. (3.2) with β_{ij} given by Eq. (3.5). Blue curves: subpopulation model with b increasing from $b = 10^{-5}$ (lightest) to $b = 1$ (darkest) with log-spaced values. Red curves: optimal SIR fits, with $b = 10^{-5}$ (lightest), $b \approx 10^{-2}$, and $b \approx 10^{-1}$ (darkest) cases. Bottom panels $\tilde{\beta}$ given by Eq. (3.4) (left, orange) and by linear regression (right, orange) as a function of b (orange), and the relative mean squared error of the true model relative to an SIR model using $\tilde{\beta}$ and N_{tot} as a function of b (blue). The plots included in this figure are from numerical simulations done with the parameters $\beta = 0.2$, $\gamma = 0.1$, $N_i = 1000$, and $K = 5$, and initial data $I_1(0) = 1$, $I_i(0) = 0 \forall\ i > 1$ and $R_i(0) = 0\,\forall i$. Both the $\tilde{\beta}$ and the regression calculations are done over an entire pandemic period (250 days).

measurement of the susceptible and infected percentage among the subpopulations, and is expected to be negative since, heuristically speaking, an increase in the number of infecteds

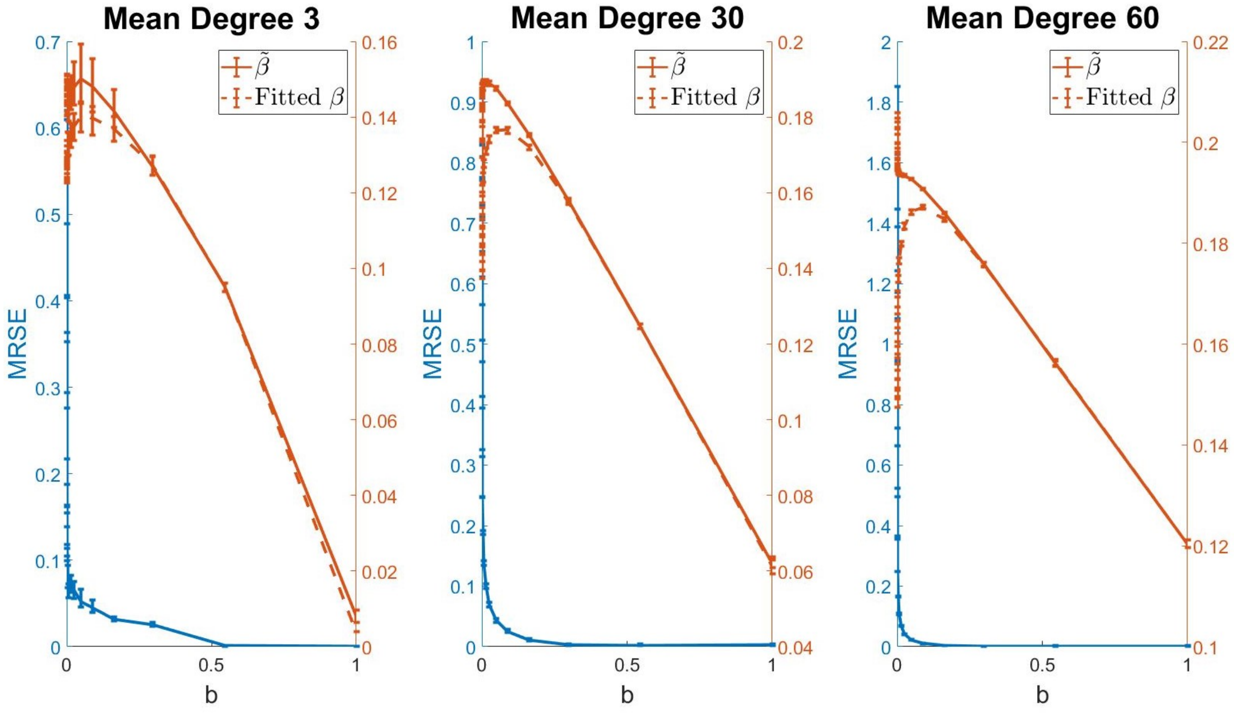


Figure 3.5: Networked subpopulations still allows for fitting by an SIR model. Erdős-Réyni networks were simulated for with three different mean degree parameterizations. 100 replica simulations were run with $\beta = 0.2$, $\gamma = 0.1$, $N_i = 1000$, and $K = 101$. Shown: Close to identical estimates are obtained for $\tilde{\beta}$ using Eq. (3.4) (orange, solid) or least squares fitting (orange, dashed). Mean Relative Squared error is shown for Eq. (3.4) (blue, right axes).

corresponds to a decrease in the number of susceptibles, at least during a surge in cases. The $\frac{\sum_{i=1}^{K}\sum_{j=1}^{K} A_{ij}S_iI_j}{SI}$ term encodes $\tilde{\beta}$'s dependence on subpopulation interaction structure. For a complete network, this term is 1, and it decreases to 0 as the network becomes more sparse.

In the complete graph case and the case with network structure, we observe that $\tilde{\beta}$ is dampened by heterogeneous community contacts. Given $\tilde{\beta}$'s correspondence with the β gleaned from data fitting, we now assume that the SIR parameter fit for β actually underestimate the true person-to-person contact rate.

3.2.4 Comparison with other estimators of disease spread

We have shown that one can generate a single population-level transmission rate, β, for a heterogeneous population. From β we may derive the basic reproduction number $R_0 = \beta/\gamma$, the expected number of secondary cases produced in a completely susceptible population, by a typical infectious individual[DHM90]. Our estimation method relies on minimizing the error between heterogeneous and an SIR model over the entire time course of a surge through Eq. (3.4). Prior estimation methods rely on the fact that in an SIR model, the parameter β describes the linearized, or initial, exponential rate of growth of the number of infectious individuals. For our model, we estimate this linearized rate of growth by two methods: 1. fitting an exponential on the first 20 days of the simulation, and 2. the next-generation matrix method from [DW02] (Fig (3.6)).

In [DW02], R_0 is computed for a general compartment model from the Jacobian matrix of the system. This matrix is evaluated at a disease-free equilibrium to determine the average number of individuals that a typical infectious person infects when the population is asymptotically disease free, i.e. $S_i = N_i$ for each i. Applying the method in [DW02] to Eq. (3.5) and making the assumption that $N_i = N_j$ for all i and j as in the simulations, we obtain:

$$R_0 = \frac{\beta}{\gamma}\Big(1 - b + \frac{b}{K}\big(1 + \rho(A_{ij})\big)\Big) \tag{3.7}$$

where ρ denotes the spectral radius of the matrix A_{ij}. In the complete network case (as in Fig (3.4)), $\rho(A_{ij}) = K - 1$, which yields $R_0 = \beta/\gamma$, matching the asymptote in Fig (3.6).

As the mean degree of our random network increases, R_0 values computed from the next generation matrix method in [DW02] and the R_0 computed from $\tilde{\beta}$ converge (Fig 3.6). On all levels of network connectedness, and mixing parameter, exponential fitting consistently underestimates the contact rate. Under-estimation results from initial slowing in early growth, due to transmission being slower between subpopulations than within them. Conversely, for

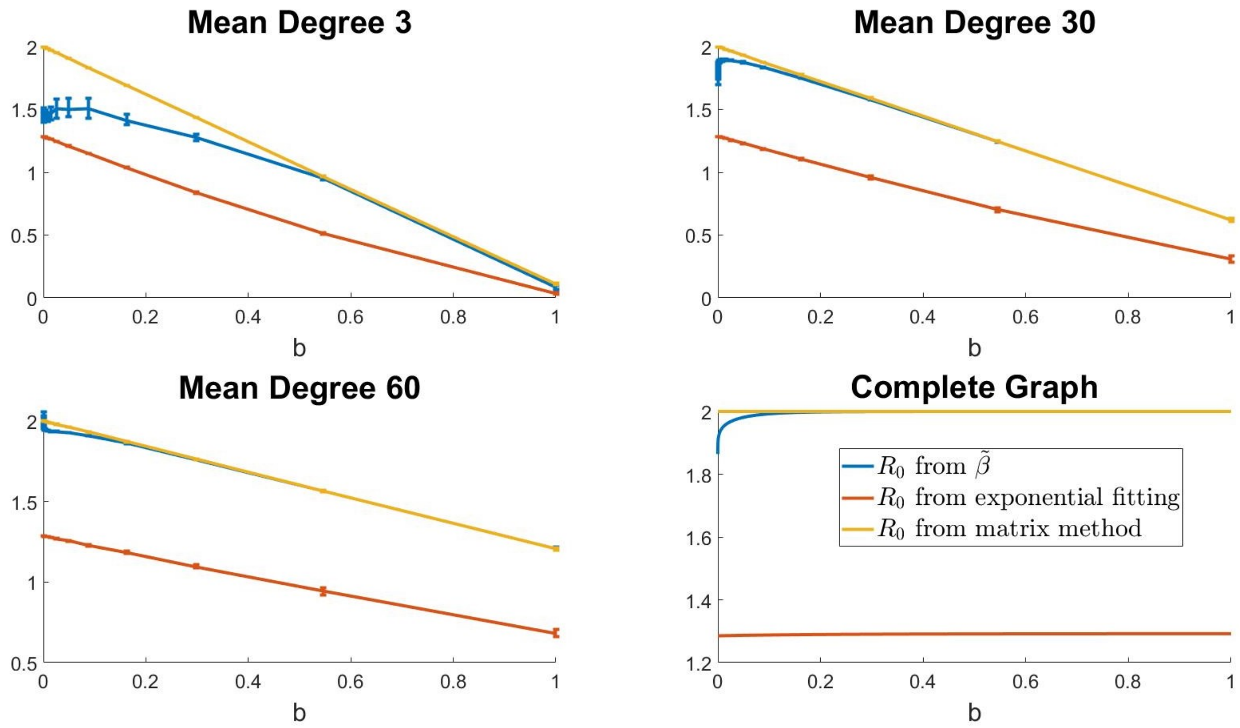


Figure 3.6: R_0 estimates from fitting entire case data curve optimally to well-mixed model (blue) agree with linearized analysis by next-generation matrix method (yellow) [DW02], but not to empirical fits to the data assuming exponential growth (red).

small values of the mixing parameter b, the next generation matrix method R_0 exceeds the optimal estimate from $\tilde{\beta}$, but the two estimators converge consistently at b values between 0.1 and 0.45, depending on the mean degree of connectedness between subcompartments.

The value of R_0 computed in [DW02] comes from the linearized dynamics; for example it provides a threshold for the stability of disease-free equilibria (see [Het00b]). By contrast the estimate for R_0 computed in this paper is computed by approximating the spread of the disease by a homogeneous model. Surprisingly, the two methods produce confluent results even under modest levels of mixing between subpopulations.

In contrast to constant-parameter growth fitting, time series fitting calculates the time-varying rate of exponential growth of the number of COVID cases, and thence infers the number of new infections caused by each COVID case. Unlike the SIR model, R_0, now called simply R, is not a constant, but typically varies over the course of a surge, and reflects not just the linearized dynamics of disease spread if an infected individual were transplanted to a population containing only susceptibles, but an actual estimate of new infections. Both SIR model and data fitting produce case curves that agree well with the surge of Omicron-variant cases, for which we use California as a representative example (section 3.2.6). However, the SIR model achieves this fit by assuming piecewise constant β, with a single (β, N_{tot}) pair covering most of the surge. We show in the next section that it is possible to forecast the end of the surge after an inflection point in a surge, and the total number of cases it will cause. Under the model, the surge ends only when the disease has been transmitted through an entire well-mixed subpopulation of size N_{tot}. By contrast, the time-varying R value inferred by Epiforecasts [Epi] gives a compelling visualization of how transmission rates decrease during the surge. However, data fitting can not distinguish between decreases in transmission rate due to inevitable decrease of number of susceptibles around each infectious individual, or due to public health orders changing the course of the epidemic. The success of the SIR model in fitting the data weights the first factor over the second, though does not discount the effectiveness of public health measures implemented at or before the start of the surge in controlling its trajectory.

3.2.5 The N_{tot} Parameter as Effective Case Surge Size

In the derivation of the SIR model, N_{tot} is the size of the population through which the disease is being transmitted. When using the model for data fitting, N_{tot} is often treated as the size of the population; e.g. state or country from which the data was sourced [BFM20]. However, real COVID cases occur in hot spots, and may not involve every individual in the studied population. In terms of our representation of this studied population by linked compartments, linkages between some compartments may be so weak that cases in one do not lead to a number of cases in the second that does not scale with the second compartment size. For this reason, we take advantage of the flexibility within our model of allowing N_{tot} be fit alongside β.

Much like how we must reconsider the fitted β in the context of heterogeneities, we must now reconsider what the fitted N_{tot} represents. The SIR model believes that N_{tot} gives the scale of the population modeled. In turn, our fitted N_{tot} parameter tells us something new and useful: the epidemic population scale that the data is conveying. Such a fitted N_{tot} can then be a new, data-driven scale by which we can estimate per-capita case numbers in the context of a given surge.

Fitting N_{tot}, however, is surprisingly tricky. In the SIR model, N_{tot} always appears coupled in a fraction with β, and as such, when the scale is not necessarily apparent from the data, only the ratio $\frac{\beta}{N_{tot}}$ can be tuned to fit. The scale of the initial conditions used to fit the data are also unreliable, in part due to a large degree of uncertainty of what the initial conditions are given cumulative case data, i.e. we have little information about how many people can be considered susceptible or removed at any given time in order to seed the fit with an accurate initial condition. Indeed, when not enough data to determine scale is present, the fits for β and N_{tot} are highly sensitive to the initial parameter estimates in our fitting algorithm (Fig 3.7).

We analyzed omicron surges (which are fit quite well by the SIR model, as explored in the next section) to determine exactly how much data is needed for scale to be fit robustly (Fig 3.7). We observe that predictive power and parameter fit robustness rapidly increase immediately after the inflection point in surge data, suggesting that the model needs to be sure of a future plateau of cumulative cases to truly determine scale independently of the ratio $\frac{\beta}{N_{tot}}$. This observation is supported mathematically by an analytical result on the

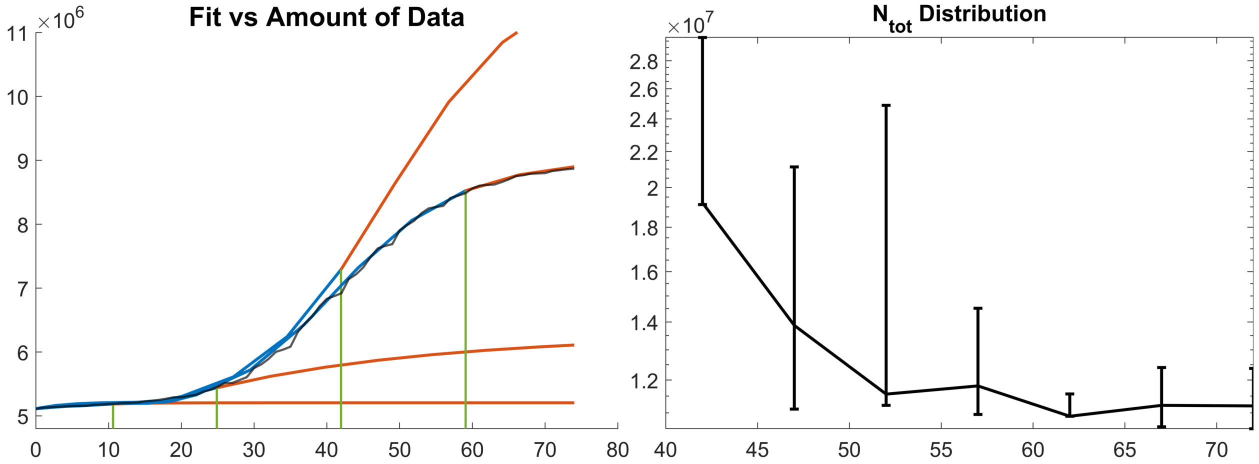


Figure 3.7: Fitting analysis of the N_{tot} parameter for California's cumulative COVID case data during the Omicron variant surge. Left: the SIR model's best prediction (orange part of curve) when given a four different portions of the entire surge data (blue part of curve). Vertical green lines indicate the point where the SIR model starts predicting. As more data is given, the model better matches the actual cumulative case data (translucent black line). Right: the distribution of fitted N_{tot} parameter values as a function of number of days of data given (horizontal axis in plot). Fit distributions are generated by taking the negative exponential of mean squared residual of the fit after randomizing the initial parameter guesses (see section SID). The distribution narrows around a more unique guess as more data is given.

eventual number of susceptibles as a function of model parameters:

$$S_\infty e^{-\frac{\beta}{\gamma N_{tot}} S_\infty} = S(0) e^{-\frac{\beta}{\gamma N_{tot}} \left(N_{tot} - R(0)\right)} \tag{3.8}$$

We see here a second indicator of scale for the SIR model present in the terminal dynamics of the model's trajectories. We reason that the inflection point in a case surge is the point after which these terminal dynamics become determined. In turn, the N_{tot} parameter reveals itself uniquely.

3.2.6 SIR Curve Fits and Parameter Conclusions from U.S. State Omicron Case Surge Data

Armed with the more nuanced interpretations of what fitted β a N_{tot} mean, we sought to draw conclusions from U.S. state data from the particularly infectious Omicron variant as it first rose to dominance among infectious individuals. Despite the relatively wide coverage of vaccines, the Omicron variant infected individuals as if they were purely susceptible, although vaccinated individuals were far less likely to end up hospitalized [MVZ21]. This high infection rate paired with the timing of two major U.S. holidays (Thanksgiving and Christmas) almost invariably led to homogeneous SIR-like dynamics for entire states (Fig 3.3), despite vast differences in population density and demographic structure. In fact, the only states not to exhibit clear homogeneous SIR dynamics appear to have extremely low-quality data. Given our early interpretations utilizing $\tilde{\beta}$ (Eq. 3.4), we conclude that some level of subpopulation interaction is happening state by state, with well-mixed dynamics occurring in the different subpopulations. Since the data considered is a surge of cases, we also expect the fitted β to represent the bulk susceptibility parameter dampened by subpopulation interaction effects due to negative covariance between susceptibility numbers and infectious numbers (Eq. 3.6). Our parameter fits for β in the following, therefore, are *lower bounds* for the true person-to-person susceptibility rate rather than actual estimates.

We chose to use California as a case study for our parameter fits due to the absolute certainty of there being at least two loosely connected population centers for COVID spread (the San Francisco area and the Los Angeles area, separated by approximately 400 miles). Much like the vast majority of states, California's Omicron surge is extremely well approximated by a simple SIR curve. The most frequently-occurring set of parameters that fit the data put the basic reproduction number at approximately $R_0 \approx 3$ (a lower bound just as β is), and $N_{tot} \approx 12$ million. The SIR model therefore thinks that California's data is actually represented by homogeneous spread among 12 million individuals. Our earlier interpretations temper this conclusion with the reality that the spread cannot be homogeneous, and leads us to the more nuanced conclusion: the spread of Omicron in California is dominated by the well-mixed disease spread amongst 1 or more distinct subpopulations totaling to approximately 12 million individuals, around 30% of California's population. Among these communities, spread has a very rapid rate of at least $R_0 = 3$. This highlights the dominant role that a closely interacting set of individuals can play in driving pandemic case trajectories.

CHAPTER 4

Flow modes provide a quantification of *Physarum* network peristalsis

4.1 Introduction

Despite the lack of a central nervous system, and the many divergences from multicellular animals, the plasmodial slime mold *P. polycephalum* is capable of complex decision making, as it explores space, locates nutrients, and links them together via a network of tubes. *P. polycephalum* plasmodia grow indeterminately, encounter and assimilate food sources, and integrate them into a network that uses materials economically while minimizing the energy costs of transport [AKF16]. The adaptive network making of *P. polycephalum* has been studied extensively, leading to insights into its ability to find paths through mazes [NG08, NYT00], solve the two-armed bandit problem [RMM16], find shortest paths in environments with variable costs of growth [BMV12], and to arrange itself optimally for delivering nutrients [DLB10, TTS10]. Additionally the tube diameters within the network, both reflect the organism's current optimization of flows relative to nutrient sources, and encode information about its previous encounters with food stimuli [KA21]. Although network morphology has attracted ample research attention, the flows within the network are relatively less well mapped. An active actomyosin cortex, enables tubes to dynamically vary their radii [Kam81]. Rhythmic contractions can create flows that span the entire network, transporting nutrients and organelles across the organism. Understanding the morphology of the network and the protoplasmic flows it creates are linked challenges, since the topology and radii of tubes with the network both sculpts the protoplasmic flows and continuously adapts to them.

As such, the study of the development of behavior and memory in this simple organism is interesting for illuminating the range of behavioral phenotypes that it may be capable of, and for understanding the mechanisms that underlie how, despite lacking a neural system or any central information processing organi, *Physarum* has a distributed capacity to assimilate and respond to information from its environment. In particular, studies have highlighted how tubes respond to external chemical cues from the organism's environment [KA21, FKW22], and to internal stimuli, including chemical signals transported by internal flows, and to the hydrodynamic stresses of the flows themselves [KAY88, AAP17].

In many filamentous fungi, protoplasmic flows are created by water uptake throughout the mycelium, which is then pushed through a network of hyphae to expanding hyphal tips, located at the periphery of the mycelium. This mechanism can produce extremely fast flows in the fastest growing fungi, including speeds of 50 μm/s, up to 100s of μm/s in the fast growing asomycete, *Neurospora crassa* [Lew05]. Multidirectional flows are possible, including flows that alternate in time within a single tube [SSR19], or that can occur simultaneously within a single hypha, when organelles are trafficked by different motor proteins, or by a combination of motor proteins and bulk protoplasmic flow [RS19]. By contrast, in *Physarum* plasmodial networks, active pumping, due to rhythmic tube contraction and dilation, propels protoplasm, and can drive it toward or away from the network periphery. Efficient transport on the network-scale may be effected by coordinating contraction phases, to create a traveling or *peristaltic* wave, in which contractions smoothly pass from each tube through its neighbors. Yet, under many conditions, contractions do not coordinate globally, creating small or local protoplasmic flows [AAP13, NYU00, MTN08].

Two classes of *Physarum* behaviors are subjects of study and modeling, and represent different time scales on which the network adapts to new information about its environment: 1. The coordinated pumping-driven flow of protoplasm [UMA86, YK84, KAY88] allowing stimulus responses on the scale of minutes [MAN17, LB11], including locomotion [RDT15, LZG15, RTU15, ZGL17], and nutrient/chemical signal proliferation [AAP17], and 2. The adaptation of network morphology, including tube diameters, densities and connectivities on time scales of 10-30 min [KA21]. Toward quantitative study of both classes of *Physarum*'s behaviors, we develop a method for measuring the ever-changing network morphology, along with a network-informed approach to mapping the distribution of contractions and dilations across this network.

Although automated analyses of sequences of microscope images allows for direct measurement of tube diameters and network morphology, and these data can be analyzed to extract summary information such as, for the network morphology, the levels of modularity or redundancy within the network [TTS10], and the phase distributions within tubes when they are contracting with the same frequencies [AAP13], the problem of mapping behaviors rapidly becomes enmired in issues of data complexity. For example, descriptors of pumping behaviors must describe time-varying radii in tens or even thousands of individual tubes, not even accounting for the additional geometric complexity created when parts of the same tube have different phases of oscillation. By contrast, when tube oscillations are coordinated across the entire network, as occurs when peristaltic waves emerge, a simple mapping of the distributions of phases may suffice [AAP13], but a range of behaviors is possible, from apparent uncoordination to global synchronization, and a focus upon phase renders opaque the pumping behaviors that occur before the emergence of the peristaltic wave, or after its dissipation. Reduced-order descriptors, in which tube oscillations are projected upon a small number of modes, can allow the structure of peristalsis, and the transition between behaviors, to be quantified. But a basis of modes needs to be detected. In [FKW22], a spectrum of peristaltic flow modes was identified using Principal Component Analysis (PCA) on sequences of pixel intensities. The PCA identified modes could be used to project pumping patterns, and synchronization of modes sensitively reports on peristalsis's first emergence and its subsequent disappearance. PCA is a broadly useful tool for identifying bases from data, but it is agnostic to the physics of the network—the geometry and connectivities of tubes—

and how these physics constrain what types of peristaltic flows the network can create. As such it is difficult to interpret how the modes identified by PCA relate to the organism's changing transportation network, or to unpack the bi-directional coupling between modes of protoplasmic flow and the morphology of the network.

Here, we identify a mapping based on network-wide flow modeling that transforms network measurements to resultant flows, effectively identifying the types of coordinated flows the network can support, and the patterns of tube oscillations that are needed to create them (called Γ, the behavior of which is sketched in Fig. 4.1). Encoded in this mapping are modes that directly identify important peristaltic contraction patterns that contribute most to a rapid response to stimuli. As a proof of concept, we study three brightfield time series (Movies S1, S3, and S5 from [KA21]) of morphologically diverse organisms over a time scale of many peristaltic patterns. We demonstrate that flow modes and contraction patterns can be followed during network evolution, and can be well represented using our identified bases. In particular, the emergence of globally coordinated modes of contraction can be read out from the time traces of the data presented via our bases, and the times and events that lead to emergence can also be visualized using these modes as they were using previous PCA analysis. At the same time, the spectrum of modes identified by our method is directly controlled by the network geometry. Remarkably, we find that the number of significant modes is affected by the size of the network, but only little by the hierarchy of tube radii within the network. The degree of isotropy of our measured networks may issue from the relative uniformity of the environments in which the networks, studied here, were imaged. Isotropic and relatively uniform networks may themselves be adaptive under conditions where growth is not biased toward or away from sparsely distributed food sources or stressors, allowing the network to respond to foods or stresses encountered from any direction.

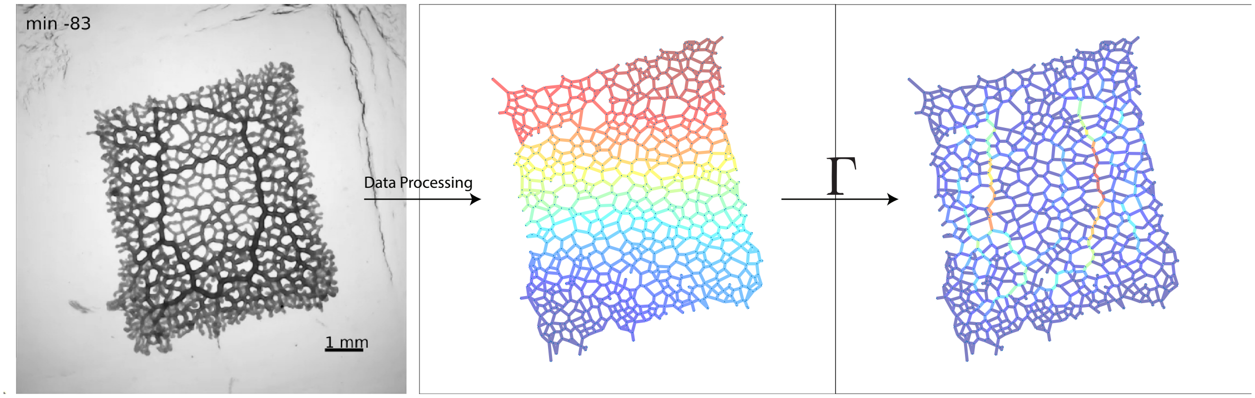


Figure 4.1: A schematic of the processing pipeline in our method. Shown is taking the first frame of Movie S1, extracting graph morphology, and finding the mode of top importance for our mapping Γ. Γ takes in a vector of tube contractions or expansions (visualized on a graph in the 2nd panel), and outputs a vector of flows (visualized in the 3rd panel). In these visualizations, red corresponds to tube expansion in panel 2 and high flow in panel 3; blue corresponds to tube contraction in panel 2, and low flow in panel 3. The mapping itself generates the modes in both panel 2 and panel 3 once data has been processed.

4.2 Methods

4.2.1 Mathematical Background

In the following, we develop the mathematics behind idealizing flows in a *P. polycephalum* organism as flows on a mathematical network.

4.2.1.1 Flow in an Expanding and Contracting Tube

The rate of flow of of homogeneous, incompressible fluid driven by pressure drops through a cylindrical tube is with the Hagen Poiseuille Law, so long as the Reynolds number $\frac{\rho u L}{\mu}$ and the Womersley number $L\big(\frac{\omega\rho}{\mu}\big)^{1/2}$are sufficiently low as previously justified for *P. polycephalum* (see [WTT11]). In these definitions, ρ and μ are the density and dynamic viscosity of the fluid respectively, L is the system length scale, u is the scale of the fluid velocity, and ω is the frequency of fluid oscillations. The Hagen Poiseuille Law states that the amount of flow Q passing through a tube is proportional to the pressure gradient Δp across the tube:

$$Q = \frac{\pi a^4}{8\mu L}\Delta p = \kappa \Delta p \tag{4.1}$$

where a and L are the radius and length of the tube respectively, and μ is the dynamic viscosity of the fluid being pushed through the tube. $\kappa = \frac{\pi a^4}{8\mu L}$ is known as the Hagen-Poiseuille conductance of the tube. This simple flow equation, however, must be modified for *Physarum* which is driven primarily by peristalsis and not by differences in pressure. Given a cylindrical tube with a prescribed, dynamic volume $V(t)$, Stokes' equations can be solved to develop a flow formula for flow that now takes into account changes in tube volume:

$$Q(z) = -\frac{\mathrm{d}V}{\mathrm{d}t}(t)\big(\frac{z}{L}\big) + \kappa\Delta p \tag{4.2}$$

where $z \in [-L/2, L/2]$ describes the longitudinal axis, with $z = 0$ measuring the center of the tube.

4.2.1.2 Conservation of Mass on a Network

We model *Physarum* as a dynamic network G with nodes $\mathcal{V}$ and edges $\mathcal{E}$. We define $\langle i,j\rangle \in \mathcal{E}$ to be the edge connecting node i and node j, $K = \big(\kappa_{ij}\big)$ to be a matrix containing the Hagen-Poiseuille conductances in edges $\langle i,j\rangle$, p_i to be the pressure at node i, and V_{ij} to be the volume of edge $\langle i,j\rangle$. By convention K_{ij} and V_{ij} are zero if $\langle i,j\rangle \notin \mathcal{E}$. We ensure that fluid is neither created nor destroyed by imposing conservation of mass at each node. In this case, this amounts to asserting that total flow into a node is balanced by total flow out. Using Eq. (4.2) measured at $z = -L/2$, conservation of mass results in:

$$\sum_{j\in\mathcal{V}} K_{ij}(p_i - p_j) = -\sum_{j\in\mathcal{V}} \frac{1}{2}\frac{\mathrm{d}V_{ij}}{\mathrm{d}t} \tag{4.3}$$

This equation is written compactly in vector form as $L_K\mathbf{p} = \mathbf{b}$ where L_K is the weighted graph Laplacian with weights prescribed by K, $\mathbf{p} = \big(p_i\big)$, and $b_i = -\sum_{j\in\mathcal{V}} \frac{1}{2}\frac{\mathrm{d}V_{ij}}{\mathrm{d}t}$. This equation can be solved uniquely so long as the graph is connected and a pressure gauge node is prescribed [FWD18]. Intuitively, such a gauge node is needed because flows are determined by pressure differences only, and so the problem must be supplied with a node to call "gauge pressure," i.e. the node that defines what all other pressures are compared to.

4.2.1.3 Relating Observed Volume Changes to Flows

Section 4.2.1.2 gives us the tools to relate observed volume changes to actual flows on the network. We define Q_{ij} to be the flow in edge $\langle i,j\rangle$ measured at the center of the tube, and define the flow to be positive if it goes from node i to node j with $j > i$, and negative otherwise. The flow in each tube measured at the center (i.e. $z = 0$) is linearly related to the pressure gradient across each tube, which in turn is linearly related to the volume change, so a linear relationship between flows and tube volume changes can be calculated. With this definition, $\mathbf{Q}$ and $\mathbf{V}$ are vectors containing the flows and volumes of all edges. We may thus derive the linear relationship

$$\mathbf{Q} = \Gamma\frac{\mathrm{d}\mathbf{V}}{\mathrm{d}t} \tag{4.4}$$

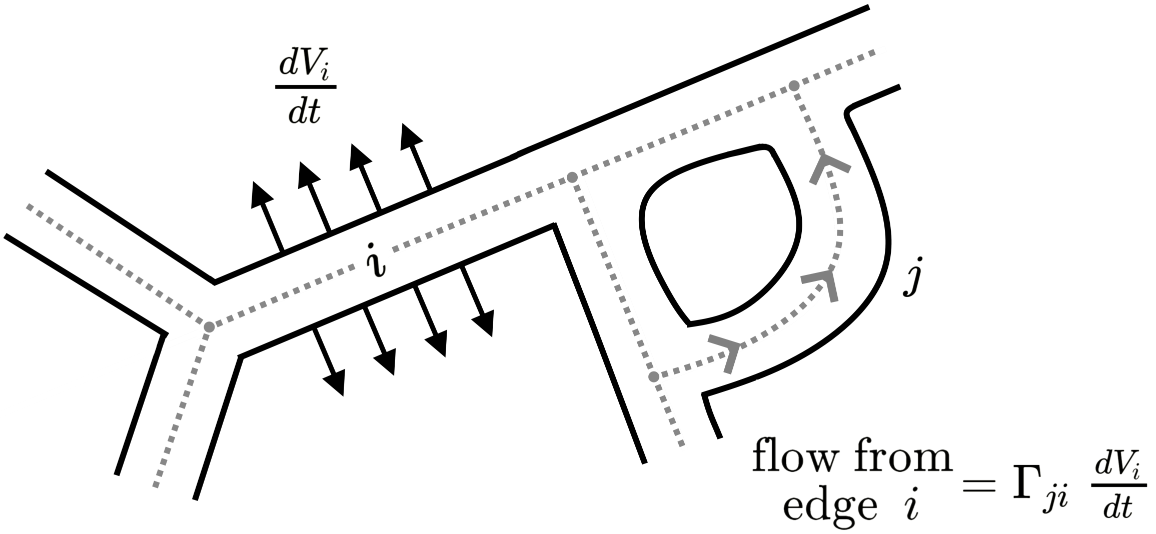


Figure 4.2: A diagram of how the mapping Γ operates on the network. The quantity $\Gamma_{ji}\frac{\mathrm{d}V_i}{\mathrm{d}t}$ represents the contribution of edge i's expansion or contraction to the flow in edge j. Edge j's flow is the sum of contributions from all other edges in the network, i.e. $Q_j = \sum_i \Gamma_{ji}\frac{\mathrm{d}V_i}{\mathrm{d}t}$.

Γ, the linear map above, is a function both of tube geometry and network topology. Mathematically, Γ may be written in terms of previously-defined quantities as

$$\Gamma_{ij} = -\frac{1}{2}K_{ij}(e_i - e_j)^T L_K^{-1} B \tag{4.5}$$

where e_i is the ith euclidean basis vector and B is a matrix such that B_{ik} is 1 if node i is an endpoint of edge k, and 0 otherwise. At a given instant in time Γ encodes how the network's observed characteristics—tube expansion/contraction, network topology—transform into the flows that transport fluid through the organism. Γ is a dense matrix, which in this case means that a volume change in any edge of the network affects the resulting flow in the rest of the network. The intuition behind how Γ produces flows from volume changes is shown in Fig. 4.2.

4.2.1.4 Analyzing Important Volume Change Modes with the Mapping Γ

Given that images of *P. polycephalum* induce highly complex networks that contain sometimes over 1,000 edges, Γ is typically a mapping that operates on extremely high-dimensional space. To make reasonable sense of what Γ is doing, it is therefore useful to capture the lower-dimensional essence of Γ. To do this, we employ the Singular Value Decomposition (SVD), which decomposes Γ into many orthogonal mappings and ranks these orthogonal mappings in order of importance by assigning each one a *singular value*. These mappings are referred to in this paper as *component mappings*. The SVD decomposes Γ as follows:

$$\Gamma = U \Sigma V^T \tag{4.6}$$

where Σ is a diagonal matrix comprised of nonnegative singular values σ_i, which are typically sorted in descending order from the top left to bottom right of the matrix, and U and V are orthonormal matrices whose columns contain the left and right singular vectors of Γ respectively. The decomposition is such that, given the ith columns u_i and v_i of U and V, $\Gamma v_i = \sigma_i u_i$. The SVD is similar to diagonalization in that important modes can be analyzed, but it has the added benefit that all numbers including singular values are real, so important component mappings can be visualized simply.

The SVD can be used to lower the dimension of a representation of a given matrix by truncating the amount of mappings one wishes to consider. For example, the mapping Γ could be (rather poorly) approximated using only 10 degrees of freedom by cutting off all but the first 10 columns of U and V and taking the upper-left 10x10 submatrix of Σ, then re-multiplying the decomposition. This idea is used extensively in Principal Component Analysis (PCA), but PCA is used for symmetric matrices whose inputs and outputs are assumed to be from the same space. In contrast, our inputs are networks contractions and outputs flows, so PCA can't be directly used.

Importantly, the SVD also identifies important modes that the mapping acts on. In our case, we use the SVD of the mapping Γ to consider the effects of geometry and topology on flow generation given primary volume change modes.

4.2.2 Extracting Network Information from Data

Here we outline the steps taken by our custom MATLAB code to extract dynamic morphological information from data.

4.2.2.1 Creating the Master Network for the Video

Over the course of a time sequence of microscope images, some tubes disappear due to pruning, and some disappear and reappear as they lose contrast with the background via contraction. For these reason, the network's topology changes with time. To account for this, we first need to create a master network that contains all possible tubes that exist over the course of the time sequence. We then define the time-varying networks as subgraphs of this master network. Here it is important to note that our process assumes that the organism doesn't move or change shape with time, which restricts the data we can process to networks that are not migrating.

To create this master network, a we take the median of approximately 50 frames exhibiting dense versions of the network to capture an artificial image where contrast between tube and background is ideal for thresholding and noise is dampened. We then create an image mask via adaptive local thresholding (MATLAB ADAPTTHRES). This mask is then skeletonized to create an image mask whose tubes are only a single pixel thick. Through a custom algorithm similar to the often-used anaskel [Fet22] called skel2graph [Wil22], the network topology (i.e. the nodes and the edges of the network) is then extracted from the skeleton. Alongside this network topology, our algorithm also returns a labeled binary image that records the actual geometric shape of each edge for use later in measuring data features.

4.2.2.2 Measuring Geometric and Topological Features for Each Microscope Frame

Armed with the master network, we now proceed to measuring frame-by-frame information. For each frame, we again create an image mask with adaptive thresholding. Adaptive thresholding considers a window of pixels around the pixel of interest and forms a threshold to separate foreground and background from grayscale data. We then directly associate each pixel in the mask with an edge from the master network. If an edge has no mask pixels associated with it, it is removed from the network for the given frame.

For each associated edge, we then compute tube volume. In previous works [BKA17, AAP13, FKW22], tube volume was argued to be inversely correlated to pixel intensity by the Beer Lambert Law, and as such intensity values of each pixels were used as an indirect measurement of tube contraction state. In our application, we utilized a compressed .mp4 video format taken directly from the supplementary materials of [KA21], yielding frames that were lower in dynamic range and resolution than previous work. Intensity measurements on the graph skeleton, therefore, suffered from a low signal-to-noise ratio, and didn't prove very useful as a proxy for volume. To smooth out this noise, we integrated intensities over entire tubes using the following model:

$$\text{Volume} = \sum_{i \in \text{tube}} DA \frac{I_{0,i}}{I_i} \tag{4.7}$$

This model reflects the fact that thicker edges are darker, and that the apparent darkness of a pixel reflects how much light is absorbed or scattered in the imaging. The sum is taken over all pixels i that make up the section of the mask associated to a given edge. Here, D is a length scale relating intensity to length, A is the area of each pixel, $I_{0,i}$ is the intensity of the background around pixel i (assumed brighter than a pixel on the tube), and I_i is the intensity of pixel i. Mathematically, this is an approximate integral of a height function over an area, which produces a volume. Since we were primarily concerned with measuring volume dynamics and not accurate measures of volume, the scaling constant D was assumed to be

1 for this analysis. With this model in hand, we were able to recapitulate the measurement smoothness shown in previous works even with our compressed data.

We then calculated an average radius for each tube by assuming that tubes are perfectly cylindrical. In this case,

$$\text{radius} = \sqrt{\frac{\text{Volume}}{\pi L}} \tag{4.8}$$

where L is the arc length of each skeletonized edge.

Finally, given that our model in Eq (4.5) relies on volume change dynamics, we computed $\frac{\mathrm{d}V}{\mathrm{d}t}$. Given that we are interested primarily in expansion/contraction dynamics and not growth, we first detrended our volume data by subtracting off a moving mean of width corresponding to two typical contraction periods. We then smoothed the data via MATLAB SMOOTHDATA and took a simple finite difference to calculate $\frac{\mathrm{d}V}{\mathrm{d}t}$ in each edge.

4.2.2.3 Identifying Flow modes on Γ

Armed with frame-by-frame network features, we could then define our matrix Γ from Eq. (4.5) for each frame. To identify the important component mappings of Γ, we employed the SVD as described in section 4.2.1.4. However, given that the total volume of the network is assumed to be conserved, we first restricted our input vectors $\frac{\mathrm{d}\mathbf{V}}{\mathrm{d}t}$ to sum to zero via a projection P onto the space of volume-conserving vectors. Identifying these volume-conserving modes was then achieved by taking the SVD of ΓP. These modes were then visualized as in Fig. 4.3.

4.2.3 Relating Flow Modes Frame by Frame for Dynamic Information

Given that the topology of the network is not assumed to be constant from frame to frame, and that the ordering of important modes is also likely to change over time, we devised a way to register similar modes to each other between frames. First, to account for differing network topologies, we only compared modes using edges common between frames via a euclidean distance. Using these calculated mode similarities between two frames, we then solved the linear assignment problem [DK01] to register modes to one another frame-by-frame. Due to the complexity of the problem, we restricted this dynamic analysis to the top 20 modes of the mapping. This assignment allowed us to form the continuous mode coefficient plots visualized in Fig. 4.6.

4.3 Results

First, we derived a physical model that incorporated changing tube contraction states into flow calculations. This model resulted in a linear map, Γ (Eq. (4.5)), that takes instantaneous contraction rates as input and outputs network-wide cytoplasmic flow. The modes and singular values analyzed in the following are those of this mapping Γ, and represent a whole-body analysis of the connection between network morphology—encoded in Γ—and dynamics. Further details on Γ are included in section 4.2.

We analyzed three sequences of *Physarum* images (Movies S1, S3, and S5 from [KA21]), which were chosen as data because they each demonstrate qualitatively different morphologies and behaviors over the course of the data. Movie S1 is a topologically consistent, densely connected network for the whole video. Movie S3 is topologically consistent as well, but with much less density than Movie S1. Movie S5 demonstrates a marked change in topology after a food stimulus is introduced, and is also fairly densely connected until pruning thins the organism. In the following, we identify morphological and dynamic measurements of these three videos and compare them.

4.3.1 Topological and Geometric Measurments

To identify the dynamical function of network morphology, we utilized the singular value decomposition (SVD) of the modeled mapping (Γ from Eq. (4.5)) that takes as input network contraction dynamics and outputs cytoplasmic flows through the network. The Γ matrix encodes both network geometry and topology via inclusion of the weighted graph Laplacian, and so an analysis of Γ is an analysis of the function of the particular network morphology, in determining how many flow modes a single network can support by tube contractions and dilations. The SVD of Γ ranks important contraction distributions by the relative sizes of the flows they are able to produce in the network.

What do these contraction distributions look like? In [AAP13], whole-network peristaltic waves were shown to be optimal when there is a linear gradient of phase across the length of the organism. Our model affirms that linear variations in contraction velocities produce largest flows. In Fig. 4.3, we observe that, despite different shapes and topologies, the highest ranked mode of Γ takes a linear gradient of contractions (top row) and creates flow along the thickest tubes that are approximately parallel to gradient direction (bottom row). Although the patterns of contraction (modes) seem to be little influenced by the hierarchies of tube radii present within the network, the flows associated with modes of contraction (bottom row of Fig. 4.3) are channeled substantially through the thickest vessels of the organism, affirming the importance of these high radius vessels to network-wide transport *P. polycephalum*.

The primary contribution of the largest vessels to the flows within the network has supported modeling and analytical approaches that simplify the network topology to the spanning tree that utilizes only these thick vessels [WFF16, FWD17, TTS10], neglecting the contributions of smaller vessels. Conversely, our calculated modes of tube contractions and dilations do not appear to differentiate between the largest and smallest tubes in the network, and consist of simple gradients of tube volume changes across the organism. The lack of appearance of the largest diameter tubes in the principal modes, suggests rather that approximating the network as having uniform tube radius, or even as a material continuum, may be pursued as a way to obtain relatively accurate approximations of the most efficient contraction modes, and indeed that these modes may be calculable only from the shape of the organism's boundary. Although this conclusion may seem at odds with both theoretical and experimental work [WFF16, FWD17, TTS10] emphasizing the importance of tube hierarchies, we note that hierarchies of tube radii are certainly important to calculating the flows created by the principal modes of contraction, with the largest tubes funneling flows created by the modes of contraction and dilation [KA21].

The spectrum of singular values of Γ measures the number of different modes of contraction and dilation the network can support, and is a functionally important readout of the network's shape. Specifically, we focus on how many modes are needed to describe the entire Γ mapping. Following conventions adopted in other mode-analyses, such as PCA, we determine the number of modes needed to make up 90% of the total map magnitude. Mathematically, finding this number is equivalent to taking the cumulative sum of the singular values in descending order and stopping when you reach 90% of the total singular value sum. For all three of the mapped networks, and over all time points, the percentage of singular values needed to take up 90% of the total singular value sum was between 43% and 47% of the total number, which is simply the number of edges within the network. This spectrum is surprisingly broad, again reflecting the fairly minor role played by the hierarchy of tube radii in selecting particularly favorable modes of tube contraction and dilation. This result, though largely consistent with the broad spectra that were previously identified through PCA analysis [FKW22], sounds two notes of caution: first, even if we accept that flows within the network are confined to a fairly narrow band of possible modes [FWD17], it does not automatically follow that the patterns of contraction used by the network to create these flows are similarly constrained in diversity. Second, the potential utility of modal analysis to reduce high dimensional data on the velocities of contraction of every individual tube, to low dimensional subspaces capable of completely describing every pattern of contraction that we might observe within the real network, is unlikely to be realized. In its place, our analysis emphasizes the breadth of the space needed to describe *Physarum*'s behavioral repertoire, and its potential responses to new cues from its environment, and echo [FKW22], in suggesting that this space allows for an extremely large array of behavioral responses, with no

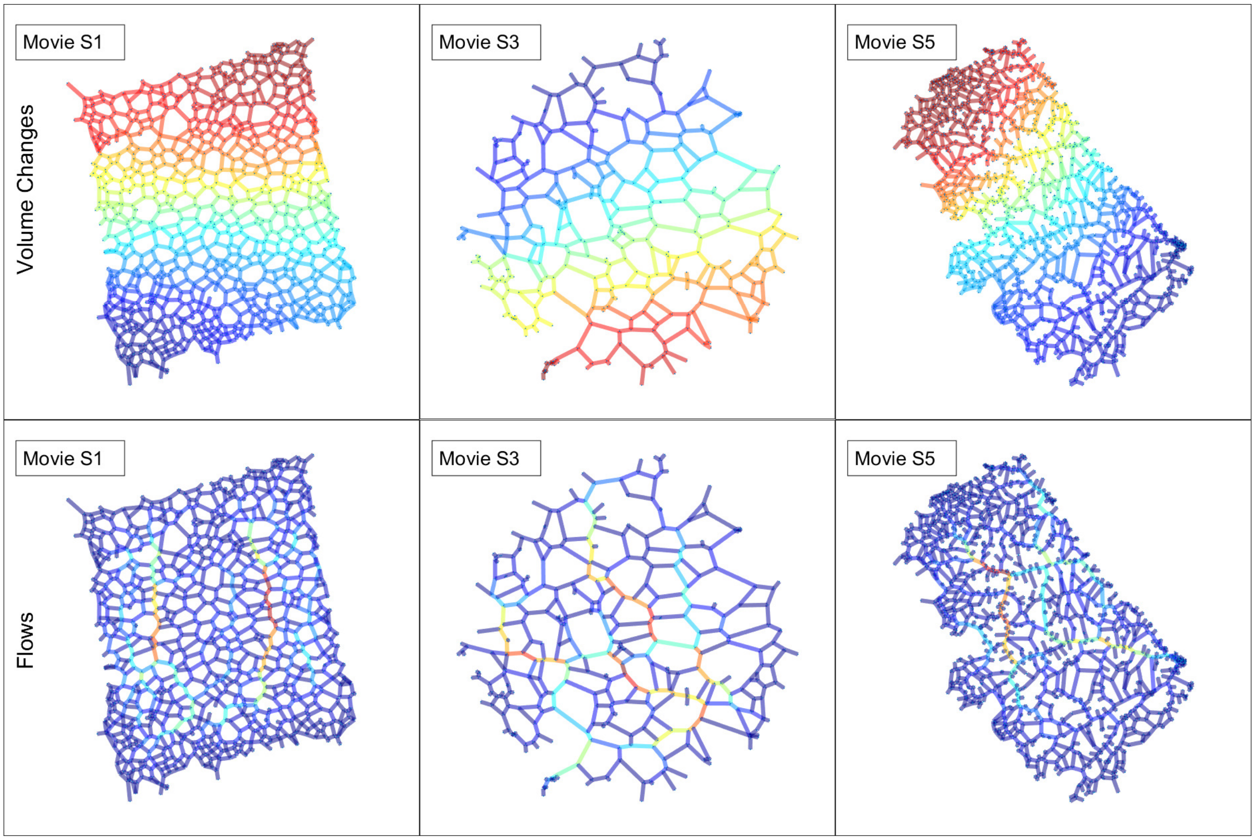


Figure 4.3: Visualized are the top component mappings of Γ for frame 1 of all 3 of our datasets. Red indicates tube expansion or high flow, blue represents tube contraction or low flow. Top row: the right singular vectors of Γ, corresponding to contraction/expansion patterns that lend themselves to produce the highest flow. Bottom row: the absolute values of the left singular vectors of Γ, corresponding to the distribution of flows created by the particular contraction/expansion patterns above them. Tubes that get the most flow (colors that are not deep blue) correspond to thicker tubes in the dataset.

clear distinction between the important and the unimportant.

To further understand *Physarum*'s spectrum of singular modes, and how it varies over time and between different *Physarum* individuals, we plot the singular values themselves, in descending order for a variety of conditions. We first plot the spectrum as a function of the absolute mode index, a number that ranges from 1 to the number of edges in the network (Fig. 4.4). The blue areas enclose the upper and lower quantile of the singular values for the entire time series, sampled every ten frames. Consistent with our analysis of the number of significant singular values, we found the spectrum to be extremely broad, with the singular value of mode n best fit by a power law: $n^{-2/3}$. Since the matrix Γ is new to this study, ongoing work is necessary to find how the real *Physarum* networks compare with other model networks. However, Γ is partly based upon the inverse Laplacian, the spectrum of which has been characterized for many model networks. Our spectrum is much broader than for empirical scale free networks for which [MM08] measures that the eigenvalues of the inverse Laplacian scale approximately like n^{-3}. They are even broader than the spectrum of Laplacian eigenvalues for a totally homogeneous network. In the limit of large networks, we would expect this spectrum to approach to the spectrum of the inverse Laplace operator on 2D domain. Absent strong asymmetries of shape, the Laplace operator has eigenvalues $\lambda_{m,n} = \frac{4\pi^2}{L^2}(l^2 + m^2)$, where l, m are the wave numbers in two orthogonal directions. Hence, the inverse Laplacian for a uniform network has $O(N^2)$ eigenvalues that are larger than $1/N$, and, correspondingly, we would expect that the nth eigenvalue of the operator would be $\sim n^{-1/2}$, which decays only a little slower than our measured spectrum. Further analysis is needed to disentangle how much of the difference is due to the structure of our operator, and how much of the difference is due to structural features of *Physarum* networks.

To probe the influence of the network structure, we first compare the spectra of real *Physarum* with synthetic networks in which we permute tube radii across the entire network, leaving the total variance in tube radii unaffected, but randomly disconnecting high conductance pathways and thus any spanning trees that the network may contain (Fig. 4.4, red regions: upper and lower quantiles of the ordered singular values). The blue and red regions show considerable overlap, indicating that the singular values are mostly unaffected by a shuffling of radii throughout the network. This overlap suggests that the specific arrangement of high conductance tubes within the network does not influence the structure of its most effective patterns of contraction, and nor does it strongly affect the scale of flows that these most effective modes can create within the network. Instead, the data suggest that the network is mostly isotropic. Given that most protoplasmic flow is predicted to be contained in the thickest tubes, as indicated in Fig. 4.3, isotropy of the thickest tubes may allow the network to push fluid in any or all directions when it encounters a new food source.

Although the precise arrangement of high conductance tubes is little distinguishable from isotropic arrangements, the presence of tube radius hierarchies strongly affects the magnitudes of flows created within the network. To probe the contribution of tube radii, we compare real networks with a second class of synthetic network, in which all tubes are uniformly given the same radius (black shaded curves in Fig. 4.4). Uniform networks are markedly lower than the correspondent hierarchical networks for the first few hundred modes, about 400 for Movie S5 and 250 for Movie S3. Singular values of Γ correspond directly to the amount of flow that a given unit network contraction can create, and so lower singular values in the black region indicate that networks with uniform thickness can produce less flow for a given contraction pattern than those with radius hierarchy. The continuous spectrum of modes, too, create flows in a variety of different direction. Combined with the importance of radius hierarchy, we observe that *Physarum*'s morphology is optimized to produce high flows in any direction.

We also study how the spectrum of singular values varies with time. Movie S5 shows a large organism that prunes itself drastically over the course of the dataset after food is introduced. During this drastic change in topology, we observe self-similar changes in the spectrum of singular values. Shown in the top left of Fig. 4.5 in blue is the cumulative sum of the singular values in descending order for every 10 frames of the video. As edges are pruned

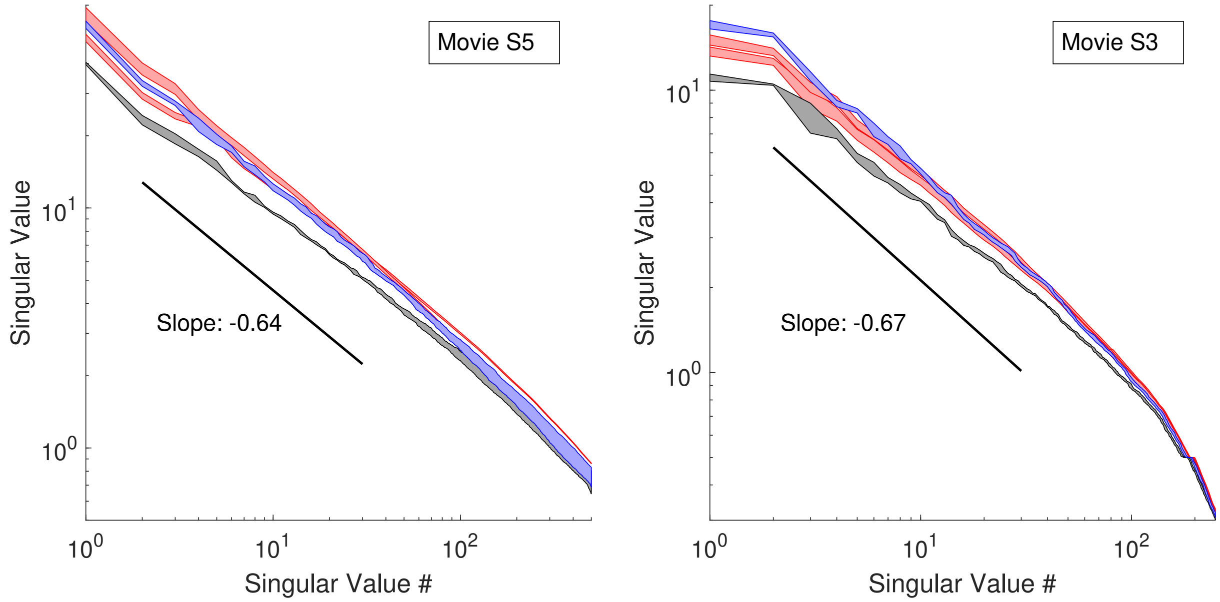


Figure 4.4: Singular values are plotted in descending order for a variety of conditions. The blue shape is the interquartile range of singular values for every 10 frames in both datasets. The black shape is the interquartile range of singular values (measured every 10 frames) when all radii in the network are set equal to their mean. The red shapes are the interquartile ranges of the singular values of synthetic networks whose radii have been randomly permuted 100 times both before and after food is added, hence the reason two seemingly distinct red shapes appear. Left: red regions are taken from synthetic networks coming from permutations of radii in frames 761-821 and 2471-2531 of Movie S5, sampled every ten frames. Right: red regions are taken from synthetic networks coming from permutations of radii in frames 461-491, 511-531, and 2871-2931 of Movie S3, sampled every ten frames.

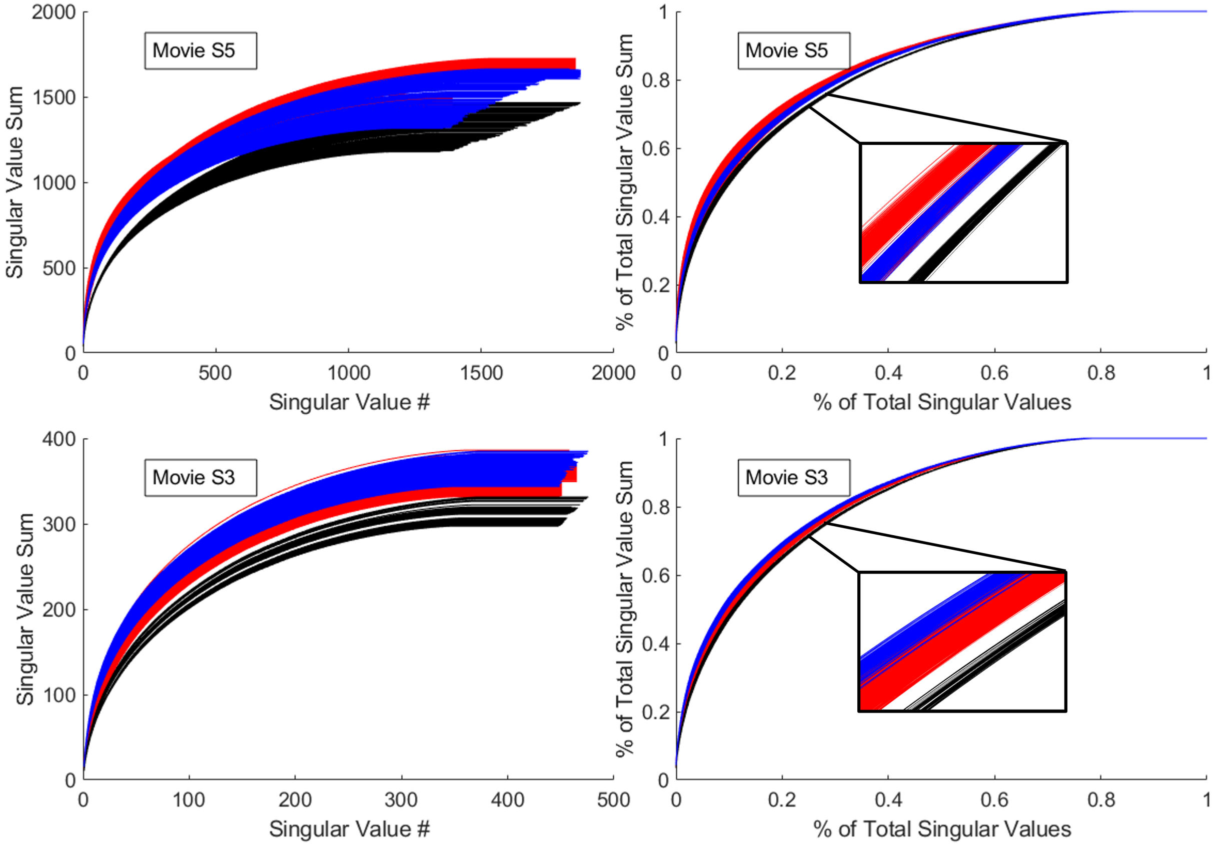


Figure 4.5: Singular value distributions, both scaled and unscaled, plotted for Movies S3 and S5. Red lines: synthetic graphs with radii of edges randomly permuted 100 times, both before and after food. Black lines: synthetic graphs whose radii are all set to the average radii for the given frame. Blue lines: actual singular value distributions for measured radii. For blue and black lines, data is measured every ten frames for the entirety of the dataset. Top row: red lines are taken from permutations of radii in frames 761-821 and 2471-2531 of Movie S5, sampled every ten frames. Bottom row: red lines are taken from permutations of radii in frames 461-491, 511-531, and 2871-2931 of Movie S3, sampled every ten frames. Left column: cumulative singular value sum of singular values in descending order is plotted versus singular value number. Right column: same as the left column, except the x and y axis have been scaled down by number of singular values and total singular value sum, respectively.

the total number of modes goes down (total number of modes is equal to total number of edges), as does the total sum of singular values. However, if the cumulative sum of singular values is normalized by the total sum, and the modes as a fraction of the total modes (Fig. 4.5, top right), the curves collapse into a single blue band. An alternative framing of this result is that in the starting and in the pruned networks, the relative contributions of any percentage of modes to the total cumulative value sum remain constant. This is not an obvious result and further analysis of Γ may elucidate why its spectrum is invariant in this way, to pruning. Pruning has hitherto been assumed to be a coarsening process, that grows the largest tubes at the cost of the smallest. Described in this way, we would expect it to emphasize stronger modes and suppress weaker modes, changing the shape of the spectrum.

We see the same self-similarity in the changing singular value spectrum over time for the entirety of Movie S3 (Fig. 4.5, bottom row). The plasmodial network in Movie S3 does not experience significant tube pruning, but it does demonstrate a significant increase in visible cross-organism peristalsis. That the mode spectrum in both organisms changes self-similarly, may reflect constraints in how strongly the singular values of Γ can be localized in a small number of modes, further supporting the inferences made before, that the specific arrangements of tubes and their radii have little effect on the singular modes. We can not rule out, however, that tube hierarchy has undetected importance, and that the invariance of the spectrum of singular values under pruning and other network changes, reflects some symmetry of how pruning and other network changes operate on different scales of tubes.

The synthetic networks created for Fig. 4.5 (A,C,red), by permuting the tube conductances in real networks, collapse to the same master curve as real networks after normalization rescaling in both the network shown in Movie S5 (Fig. 4.5B), and the network shown in Movie S3 (Fig. 4.5D). The spectra at different time points overlap for real and synthetic networks over most of the significant singular values, and though the real network has a slightly broader spectrum than the synthetic network in S5, the places are reversed in S3, suggesting that there is no significant difference between the normalized spectra. Conversely, in synthetic networks with uniform radii (black lines in all four panels of Fig. 4.5), the spectrum lies slightly underneath both real and permuted spectra, both with and without normalization. This small difference, consistent between the two networks suggests that the singular values of the top modes of the uniform network are not as top-heavy, suggesting that mild selection occurs for the first few modes in networks with some radius hierarchy.

4.3.2 Measuring emergence of coherent pumping

Spectral analysis of Γ gives us a readout of what internal flows a network can generate through tube radius changes. When real measured volume changes are projected onto these modes, they provide a reduced basis for describing the pumping that occurs in the network. A complication here is that, as shown in section 4.2.2, the topology of the measured network changes frame-by-frame, albeit by the gain or loss of only a handful of edges among hundreds for the networks analyzed here. Because of network changes, Γ's matrix representation changes size over time. However, although the total number of contraction modes changes over time, we find that the most important calculated contraction modes maintain their order, and approximately, their singular values. We proceeded, therefore, to analyze the coefficients of the projections of the data onto the top 20 modes of the system (see section 4.2.3) during intervals in which we visually identified the emergence of coherent tube oscillations. Movies S1 and S3 were chosen because they contain sections of video centered around both the introduction of food and also visual peristaltic waves moving through the organism. The coefficients of the first modes are shown for excerpts of both video sequences in Fig. 4.6.

How does the presence or absence of visual peristalsis affect the readout of our mode coefficients? Movie S3 probes a rather stable, circular network with barely perceptible contraction dynamics before food is added. At frame 500 (minute 0), food is added outside of the frame, and coordinated peristaltic contraction patterns become visible. Movie S3's mode coefficients (Fig. 4.6, left) at the moment of food introduction show clean oscillatory

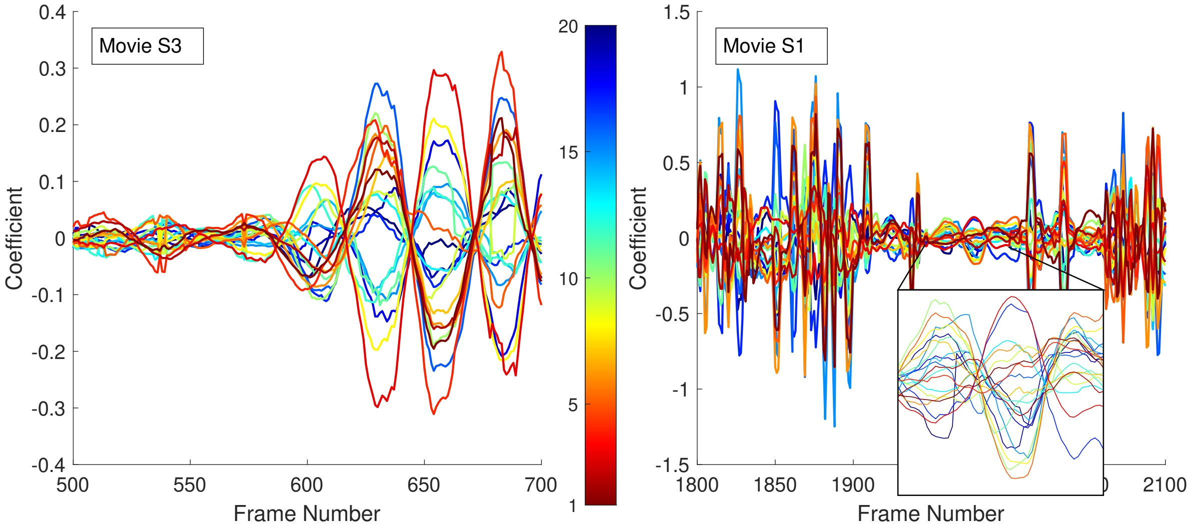


Figure 4.6: The top 20 mode coefficients are visualized for Movies S1 and S3. Colorbar shows ordering of modes from first (red) to 20th (blue). Left: Movie S3's coefficients form coherent oscillations throughout, increasing in amplitude over the excerpted frames. Lower modes attain higher amplitudes. Right: mode coefficients in Movie S1 are only intermittently coherent. The data corresponding to periods in the video without clear and visible peristaltic contraction patterns do not seem to show oscillation. The long period of coherent oscillation corresponds to a visible peristaltic wave in the video.

behavior in line with the previously measured peristaltic frequency of about 90 seconds (30 frames) [AAP13]. We see the mode coefficients reflecting the visual emergence of coordinated peristalsis: around 240s (80 frames) following the introduction of the cue the amplitude of all coefficients increases X-fold. Amplitude increases are greatest for the top modes, showing that pumping occurs preferentially in the modes with greatest flow creation potential (i.e. greatest singular values).

Movie S1's mode coefficients (Fig. 4.6, right) show the emergence of coherent pumping from initially highly disordered contractions. Visually in the large, dense network imaged in Movie S1, contraction patterns appear random across the organism. There are brief instances, however, when a cross-organism peristaltic wave presents itself—here we see one such instant at around minute -4 (around frame 1900) of the video. Before this instant, we see large-amplitude, unaligned oscillations on the order of 10 to 15 frames which could likely come from a mixture of uncoordinated pumping and aliased (noisy) data. When organism-wide pumping is observed, we observe clear, coherent oscillations whose period (90 seconds) matches published data on *Physarum* time-oscillations. It is as of yet unclear what stimulus, if any, triggered the appearance of coherent contractions, as food had not yet been added, nor what triggered their disappearance.

Results from our modal analysis are qualitatively consistent with previous descriptions of *Physarum* behavior. Specifically, we find that in networks presented with no food source, tube wall oscillations may be coherent, with small amplitudes (S3), or incoherent, likely presenting as high frequency modes due to a combination of noise and aliasing effects (S1). When presented a food source, we see existing coherent multimodal oscillations gain amplitude. Coherent multimodal oscillations emerge from noise and disappear spontaneously. We obtain additional quantitative data on the time delays between the food source being presented, and the amplification of coherent oscillations, as well as on the suddenness of amplitude increase.

Acknowledgements

This work was supported by a Research Grant from HFSP (Ref.-No: RGP0001/2021)

CHAPTER 5

Conclusion and Discussion

In this dissertation, I endeavored to explain the complex phenomena of COVID-19 dynamics and *Physarum* behavior using interpretable, physical models. We began by comparing COVID-19 cases between populations, which is an important part of building narratives for understanding the impacts of the disease, and the roles of policy, inequality and racism in shaping its severity. But case numbers can not be compared without correcting for differences in population sizes. Normalizing by the size of the population can lead to misleading comparisons, if COVID-19 incidence varies in systematic ways among individuals in the population. Here we sought to make comparisons between populations (states), under the simplifying assumption that each contains an effectively well-mixed subpopulation of COVID-19 contacting individuals. Importantly, detecting the size of this population is an element in our algorithm.

Clustering of data from different states illuminates a second, general, question, that was explored more deeply in Chapter 3 of this work. Specifically, although modern tools in epidemic modeling can incorporate population-heterogeneity, because of their many parameters, these models are challenging to fit to real data from the COVID-19 pandemic. The assumption that cases are dominated by transmission within a subpopulation of COVID-19 contacting individuals, who are effectively well-mixed, motivates us to further consider that there may be a limited number of archetypal case growth curves among this subpopulation. We find through clustering, that when properly normalized, and in spite of very different geographies and demographic profiles, cases in different states follow one of remarkably few

(4-6) growth archetypes.

Some of the conclusions that follow from the clustering are expected—for example we found that the late stage of the pandemic behaves distinctly from the early stage, as indicated by the fact that the clustering algorithm separated the two phases without direction. Due to the wildly different social and epidemiological conditions between the two time periods, this is unsurprising.

The first phase conforms extremely closely to SIR dynamics, as indicated by Fig 2.3, suggesting an effective mean-field description of the early dynamics. The fits shown in the figure come from least-squares model smoothing: we found appropriate constant transmission rate β and population size N values for the SIR model that best fit the data in a least-squares sense (see section B.1). Importantly, fitting N to the data enables the entire first wave to be fit with constant disease transmission parameters, and thus with a constant basic reproductive rate R_0 [DHM90]. In addition to their greater simplicity, the constant β, N models draw a very different conclusion about how the wave ends than a time-varying R_0 model—for example the fits for Washington (shown in Fig 2.3) give an R_0 value of 1.14 within an affected population of about 27000 . At this value of R_0, the disease spreads through its affected population, and case numbers only start to decrease once $1 - 1/R_0 \approx 12\%$ of the affected population has acquired resistance to the disease. The actual downturn occurred well before this 12%. The fits therefore are consistent with almost immediate success of shelter-in-place measures in preventing the disease from spreading beyond a fixed number of closely-contacting individuals, but continued, unchecked spread of the disease within this group. This is in contrast to other, more complex methods of data fitting that relied on a time-variable reproduction number R [AHT20, TSv19, CFF13, AHB21]. Hence, although actual transmission behavior is certainly heterogeneous and occurs in multiple subpopulations, a simple, constant, bulk parameter suffices to describe early epidemic dynamics within populations with a large percentage of susceptible individuals, as explored further in chapter 4.

What reveals itself unexpectedly is just how similar the late stage dynamics are from cluster to cluster. There are three factors separating late stage surge dynamics of cumulative cases on the log scale indicated by Fig 2.4: One; the overall rate of increase, i.e. the overall "slope" of the curve; Two; the degree to which the "S" curve deviates from the straight line; Three; the time scale over which the curve displays its "S." A possible fourth factor is the precise position of the inflection point of the "S" curve. This is significant since, despite wildly varying conditions between and within states, there seems to exist only three or four bulk parameters which suffice to describe the spread of COVID during a surge. This leads to a reassuring discovery: during surges, the simplest models suffice. At the beginning, the SIR model does well to describe dynamics, and later, some new but equally low-dimensional model can likely be found.

We also observe that a pair of states sharing behavior at an early phase is not predictive of the same pair sharing behavior later. This can be seen by examining the totally different cluster structure between the two phases, and is confirmed empirically using a permutation test. We measured the number of states sharing the same cluster, concluding that 37 unique pairs of states shared a cluster in both the early phase and the late phase (Texas and South Carolina, for example). This number is not significantly high as indicated by permutation test (approximately 40% of statistics from the permuted data were higher, see Table 2.3), indicating what we see at a glance: behavior shared early does not imply behavior shared late. Comparisons which were important towards the beginning of the pandemic may not necessarily be appropriate at later times, further supporting our comparison of entire time courses.

Our clustering method provides a data-driven, nonparametric way of classifying and comparing State-by-State COVID dynamics. This new approach revealed that comparisons between states using metrics that are devised *a priori* are perhaps misleading. For instance, California and Florida have often been compared [Woo, LM, Cur] since these are two states boasting large populations with multiple urban centers and warm climates, yet having very different responses to COVID—for example, during the second surge, both of the largest school districts in California (Los Angeles and San Diego Unified) were closed to in-person instruction, while in Florida a State order required that all schools offer in-person instruction after July 2020. Upon Florida's reopening in August, per capita new cases were 20 new cases per 100,000 in California and 45 new cases per 100,000 in Florida. However, our analysis emphasizes that heterogeneities in COVID case intensities across each State make the total State population a weak normalizing factor when comparing different states. When normalized by the detected size of the COVID-affected population we find that California's cases grew in line with Colorado and New Mexico, but most surprisingly also in line with Indiana, Michigan, and Vermont. Conversely, in the late stage Florida's dynamics (and Arizona's) distinguished themselves from the rest of the states, including California, mostly due to a wildly different "slope" of the general late-stage increase on the log scale, and partly due to a much earlier and drawn-out surge (Fig 2.5) (Florida's surge began in late October, and other states' tended to begin in late November, coinciding with Thanksgiving). This singles out Florida as unique from every State, not just California, and tempers any conclusions that can be drawn from the simple two-way comparison. Our results indicate that California, whose late behavior is represented in Fig 2.5 cluster 5, experiences a gentler exponential growth rate with a more pronounced downturn that happens earlier (relative to the late surge) than any downturn in Florida's dynamics. In particular, differences in curve starts and ends will cause instantaneous metrics to give misleading comparisons between the two states.

Among the variables tested, only proximity of states showed any statistically significant correlation with the clustering devised. This in itself is an important result when evaluating the probable effect of measures that may carry other public health costs, such as school closures [ADS20]. While the cluster a state belongs to controls the trajectory of the case numbers, the normalization of these numbers, which we have interpreted to be the number of COVID-contracting individuals, may be yet influenced by these or other variables, and independent evaluation of the population sizes, and whether they can be predicted from population, geographic, political, or public heath data is certainly warranted.

If the eventual goal in COVID data comparisons is to decide which health measures, public behaviors, and demographies lend themselves best to weathering a pandemic, then we must take care to compare the right data at the right times. Our analysis highlights two features. First, we show that pairwise comparisons of states can only be made in the context of a per-population case number that indicates the number of COVID-affected individuals, and not the total population of the State. Second, once the normalization is found, states conform to one of a small number of archetypal case curves. For early phase of COVID we can explicitly fit these case curves by an SIR model; in the latter phase, the well-mixed model no longer fits the data, yet the robust similarity of case curves suggests that models with comparably few degrees of freedom may yet be used to fit the data. Given the abundance of pandemic data now available, we have the tools to examine entire time courses at different scales (e.g. city or country) within the U.S. and internationally and compare in context.

In Chapter 3, our analysis of data from states and territories provide empirical evidence that there is enough mixing between the disease hotspots within each state or territory that approximation by a well-mixed model is appropriate. The concordance between β obtained by curve fitting and Eq. (3.4) affirms that least squares fitting extracts information on the subpopulation interaction structure and dynamical asymmetry between subpopulations, and highlights the SIR model as a coarse-grained model for disease transmission in heterogeneous populations. Remarkably, under moderate levels of transmission between subpopulations, our model-fitting approach, which is based on fitting the entire disease case curve, produces parameter estimates that agree well with the next generation matrix model based only on the linearized dynamics that include information only on initial epidemic features. We additionally gain the ability to reinterpret the total epidemic size N_{tot} as a new, data-driven normalization factor. We may therefore construct case rates with respect to surge size, effectively considering normalized cases only among the population actively transmitting.

In general, simplicity of fitting to real data makes SIR (and similar well-mixed models, such as SEIR) powerful tools for predicting the ongoing course of an epidemic. However, the assumptions behind the model drastically simplify real patterns of human interaction and disease transmission, and the question of estimating community interaction structure remains. Indeed, an important corollary of our analysis is that, with even modest levels of connectedness between the subpopulations, they can function as well-mixed, making the details of the substructure undetectable by a well-mixed model. Although heterogeneity-capturing models aim to render these interactions, they introduce additional parameters that often must also be fit to the data. It can be hard to distinguish improvements in fitting due to greater model realism from improvements due to increased parametric flexibility.

Data fitted analyses make no assumptions about the underlying patterns of interaction, but create short term predictions based on fitting evolving exponential growth curves to the data. SIR models do not fit every phase of the case data; for example we could not fit the second surge identified in chapter 2, likely because of the complicating effects of time evolving contact rates caused by changing work patterns and public health orders, and the presence of previously recovered individuals within the population. By contrast, although models of the the Omicron surge encounter similar complications, reluctance to reimpose social distancing measures, and the ability of Omicron to readily infect vaccinated and previously-infected individuals lead to a time course that is closer to the first surge, and that can be well-fit by an SIR model.

After analyzing how, when, and whylow-dimensional models apply to COVID data, we moved to modeling contraction dynamics of *P. polycephalum* in Chapter 4. The coordinated contractions of tubes in *P. polycephalum* enable the organism to achieve a remarkable repertoire of behaviors, including generating network-wide transport of nutrients and protoplasm, whole organism migration and network remodeling. Here, we develop a new tool for quantifying how many network-wide contraction patterns a single organism is capable of, anatomizing those patterns, and identifying, instant-by-instant, which modes are active within the network. Our method originates in a linear mapping, Γ that ties together all of *P. polycephalum*'s directly measurable quantities—topology, geometry, contraction dynamics—into an easily-interpretable and mechanistic model of fluid flow. The mapping generates theoretical contraction modes for the given network morphology that are ranked by how much flow they create and thus may be linked to notions of efficiency of transport that have already been extensively studied for the organism [GC78, AAP13, BH13, WTT11, BMV12]. Our identified flow modes correspond remarkably well to purely data-driven measurements of contraction modes. Since Γ also directly computes the flows that arise from these contraction distributions, when paired with data, Γ can be used as a measure of how much the network is working to transport fluid at any time.

The modes of Γ can be interpreted as the directions in which contractions can most readily generate flows within the network. The modes are readily computed via singular value decomposition, and the spectrum of modes reflects the range of flows that a network is able to easily support. We find that the greatest flow (largest singular value) modes align with the the optimal peristaltic phase distribution of [AAP13], being a linear gradient of contraction strengths.

Although our findings are confluent with previous qualitative and quantitative descriptions of *Physarum* behaviors, they go beyond existing analyses and methods, such as PCA. Importantly we may interpret the modes of Γ as directions in which the large internal flows may be created, allowing the contraction pattern of the network to be physically linked to its architecture. Analysis of pumping modes through the mapping Γ, highlights the role that network hierarchy plays in creating directions of greatest flow.

Surprisingly, we found the spectrum of possible modes within all of the analyzed networks to be remarkably broad; and that the number of significant singular values was not strongly affected by the hierarchy of tube radii seen within the network. This result should not be interpreted to mean that tube radii hierarchies are unimportant, since the singular values of the (real) hierarchical networks are greater than in (synthetic) networks in which all tube radii are made uniform. So hierarchies of tube radii allow larger flows to be created within the network, meaning that they increase the flow response to changes in tube radii. Yet, we found relatively small differences between the spectra of real hierarchical networks, and those of variable-tube networks in which all tube radii had been randomly permuted. This result is somewhat at odds with previous work that has stressed importance of the largest vessels in setting the flows that occur within a network, and even suggested that a spanning tree of largest tubes represents all of the peristaltic modes that the network may carry [FWD17]. By contrast our results emphasize that the spectrum of possible network pumping modes is quite isotropic, with a large number of ways of distributing strong flows through the network. These spectral features may reflect the life history stages in which experimental data was collected, in the main video sequences analyzed here, a cut *Physarum* network is imaged before and shortly after it encounters a new food source. Accordingly, the networks that are mapped here may be adapted toward isotropic vessel arrangements, since both the direction in which new food will arrive, and the directions with which nutrients will need to be transported, are not known to the network.

It would be interesting to contrast the results presented here, largely confined to initially static networks that then assimilate small numbers of food sources, with analysis of how the number of modes in the spectrum evolves when a network undergoes substantial pruning – i.e. that encounters and assimilates disparate and wide spread food sources, and where the dominant form of morphological adaptation is one of vessel loss to build an optimal linking network [TTS10]. In these networks, adaptation may be for a small number of peristaltic modes, as the network organizes itself around a known, small, set of known food sources, rather than in preparation to encounter new food sources.

We add some notes of caution about the kinds of behavioral data for which our method can provide quantification. The results in this paper are derived from compressed video files, at typical resolutions of 70-200 pixels per millimeter, and since tube diameters are on the order of .05 mm (as few as 4 pixels) to .2 mm, pixel noise in the radii of individual tubes are visible in some of the data. When spectral expansions are performed on the contraction dynamics, aliasing effects occur, in which we fail to see the expected decay of mode amplitudes across modes and in fact, there is apparently high frequency ringing in modes. Thus, although the emergence of concerted patterns of peristalsis leads to clear signals, we can not dismiss that the high frequency oscillations in Fig. 4.6 result from processing noisy measurements, rather than reflecting real patterns of organization of contractions. Careful assessment of the quality of data is important before using this, or any spectral method, since in every mode, the coupling of contractions across all tubes means that small measurement errors in the finest scales can produce contaminating signals in any mode. In further analysis we will seek to apply our model to high resolution microscope images, and to develop anti-aliasing that filter out the smallest scale contractions when projecting contraction data onto our modes.

We leave for future work the use of Γ-modal analysis on a wider variety of data, especially sequences in which the network is allowed to travel in search of new food sources. Even when networks are not migrating, quantification of memory and long term behaviors requires tracking contraction modes over longer time sequences, during which the topology of the network can change. In this study, we were able to project tube radius changes onto a small set of modes that remained relatively constant over the entire sequence of the video. Mode ordering is not in general guaranteed, when the Γ matrix varies, due to the altering plasmodial network. Alternative data assimilation method, including Multi-Hypothesis Tracking [Cor15], may allow the challenges associated with time-varying modes, to be overcome.

Appendix A

Detailed Methods for COVID Clustering

A.1 Isolating Early and Late Case Curve Phases

To identify which sections of cumulative case curves corresponded to surges, we first preprocessed data from The COVID Tracking Project [The] to estimate the instantaneous number of infected individuals in each State. Instantaneous case numbers were estimated by assuming that each new measured case was reported 5 days after actual infection occurred, and then modeling the length of time each individual was infected with a seven day delayed exponential decay curve that assumed half of infected individuals would no longer be infectious after fourteen days (Fig A.2). Then, the beginning and end of the first and final concave peak in the resulting infectious curves were recorded, and the cumulative case curve associated with these two peaks was isolated (tables A.1, A.2, and A.3). The first (early) phase occurs at the beginning of each State's outbreak without exception, and the final (late) phase in the data used occurs usually around the period of the Thanksgiving and Christmas Holiday season, sometimes extending into January of 2021.

A.2 Clustering Early and Late Case Curve Phases

To cluster our isolated phase case curves, we first devised a metric for computing dissimilarity or distance between phases. To account for differences in scale that are expected between states of different population sizes as well as to account for differences in onset of the waves due to different times of first COVID introduction to the State, or differences in the timing of public health measures such as Stay at Home orders or school closures, we measured the residual dissimilarity between each pair of curves after they had been aligned. Mathematically, this calculation amounts to defining a dissimilarity function $f(\mathbf{X}, \mathbf{Y}, T, S)$ for two time series $\mathbf{X}$ and $\mathbf{Y}$ given a shift in time T and shift in scale S

$$f(\mathbf{X}, \mathbf{Y}, T, S) = \begin{cases} \frac{1}{\ell(\mathbf{X})} \sum_{i=1}^{\ell(\mathbf{X})} \big(\log\left(X(i)\right) - (\log(Y(i-T)) + S)\big)^2 & , \quad \ell(\mathbf{X})/\ell(\mathbf{Y}) \notin (0.5, 2) \\ \infty & , \quad \text{otherwise} \end{cases} \tag{A.1}$$

where $\ell(\mathbf{X})$ is the number of time points recorded in the time series $\mathbf{X}$. The ∞ is here included so as to invalidate the comparison of two states where one State's surge length is over double the other's. The final dissimilarity metric was then constructed via minimizing this function applied to both the cumulative data and the daily data over T and S for every pair of states, i.e. if $\mathbf{X}_i$ is the ith State surge, then the dissimilarity between surge i and j is

$$d_{ij} = \min_{(T,S)\in\mathcal{T}} \Big\{ f(\mathbf{X}_i, \mathbf{X}_j) + f\big(\text{diff}(\mathbf{X}_i), \text{diff}(\mathbf{X}_j)\big) \Big\} \tag{A.2}$$

Here $\mathcal{T}$ is the set of pairs (S, T) such that T is an integer where shifting a time series to the right by T satisfies the property that 80% of the smallest curve's length overlaps with the other curve after shifting, and diff($\mathbf{X}$) is a discrete derivative;

$$\text{diff}(\mathbf{X})_i = \begin{cases} 0 & , \quad i = 1 \\ \mathbf{X}_i - \mathbf{X}_{i-1} & , \quad i > 1 \end{cases} \tag{A.3}$$

Such a minimization attempts to match two surges' values and derivatives as closely as possible.

d_{ij} was computed for every pair of U.S. states including the District of Columbia to generate a set of pairwise distances which was used to cluster. The WPGMA algorithm [Sok58] was then performed on these pairwise distances to cluster phases together, where the number of clusters was chosen heuristically by finding an "elbow" on the curve generated from plotting the number of clusters versus the cutoff distance (Fig. A.1). We then curated the clusters and manually combined suitable clusters together when such combinations maintained curve homology, and excluded two curves that did not cluster well with any other curve (Florida and Arizona late phase). Additionally s_{ij}, the minimizing S value of the function f (i.e. the upwards or downwards shift in the curves on the log scale), was stored, and is used later in section IIB as a measure of the scale difference between COVID-contracting populations of each State. Specifically, our estimated ratio of COVID-contracting populations between State i and State j is given by $e^{s_{ij}}$.

A.3 Correlating Clustering with other Variables

We look for explanatory variables than can elucidate similarities between State case curves within clusters. These variables may be either categorical (whether neighboring states or political party of the governor correlate to clustering, or whether states tended to stay clustered throughout the pandemic) or numerical (whether population centroid differences, Trump voting percentage differences, testing positivity rates, or school reopening times correlate to clustering). We use permutation tests to calculate p values against these variables explaining the observed clustering. For variables dealing with pairwise relations, we performed a modified Mantel test, which correlates our dissimilarity metric d_{ij} with the pairwise relation at hand. We restrict permutations to occur only within clusters as opposed to over the entire data set since the scales of dissimilarity scores are meaningful only in the cases where the curves being compared have similar shape, i.e. curves that are in the same cluster. For data that we did not interpret as pairwise or whose relation to our clustering was not readily measured via correlation, we devised a statistic to measure the performance of our clustering with regard to the variable in question and tested it against 50,000 permutations of the clustering IDs. The statistic tested depended on the variable at hand. To test our clustering against the political party of the governors of each state in the cluster, we first computed the percentage of governors in each cluster belonging to the most represented party in the cluster (for example, a hypothetical cluster with five Democratic governors and three Republican governors would be measured as 5/8). We then measured our statistic of the overall clustering to be the average of these percentages across all clusters. To measure whether our clustering tended to group certain states together both in the early phase and later phase of the pandemic, we computed the average percentage of pairs of states in each cluster that were clustered together in both early phase and late phase. For testing whether neighboring states tended to be clustered together, we computed the average percentage of states in each cluster who shared a neighbor in the cluster. To test whether our clustering was influenced by school reopening times or testing positivity rates, we computed the average variance of these variables over all clusters. A summary of the statistics used can be found in table II. For all of these tests, the p value reported in table II refers to the fraction of random permutations that yielded a higher statistic than that of the actual clustering.

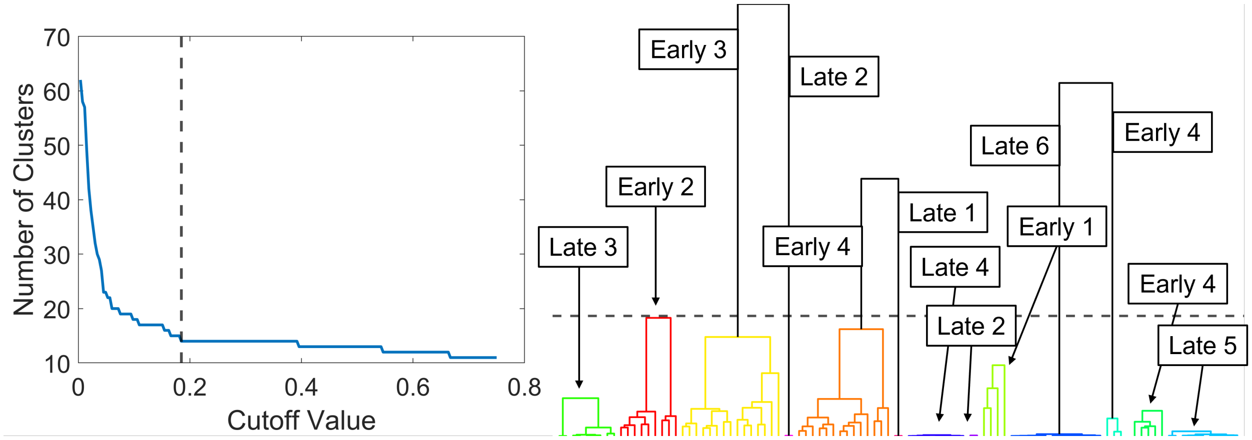


Figure A.1: Left: the number of clusters identified is plotted against the distance used by the WPGMA algorithm [Sok58] to differentiate different subtrees on the dendrogram. The vertical dotted line represents a point at which raising the cutoff value no longer decreases the number of clusters by an appreciable amount. Right: the dendrogram. The black dotted line represents the chosen cutoff value corresponding to the left pane. Different subgroups representing the 14 uncurated clusters are shown in different colors. The cluster labels both early and late are indicated with the boxes and arrows on the dendrogram.

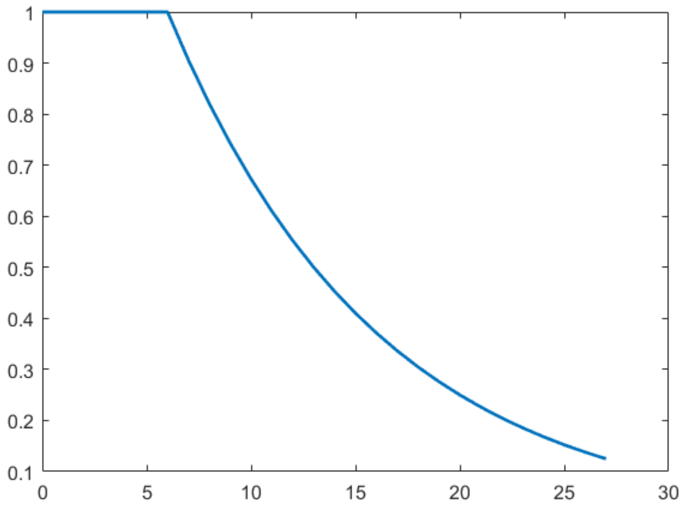


Figure A.2: The assumed remaining fraction of people who are still infected if their infection started x days ago, where x is the horizontal axis.

A.4 Estimating Current Infectious from Cumulative Case Data

In order to determine when Covid surges took place, we first needed some way of estimating how many individuals were currently infected day by day from cumulative data. To this end, we first determined the number of new cases recorded per day via differencing (described in Eq. 3 of the main paper). We then assumed that, on average, individuals remained infected for at least seven days, and thereafter this number decreased to one eighth exponentially for an additional three weeks, and then immediately to 0 (Fig. A.2). We then assumed that there was a five day reporting delay between becoming infected and testing positive, although the exact timing of this reporting delay didn't seem to affect the qualitative behavior of the resulting estimated curve, nor did adjusting the assumption of how long individuals were infected. This is a simplistic and relatively ad hoc way of determining the current infectious numbers, and is less precise than more complex methods, but has the benefit of making very few modeling assumptions beyond anecdotal observations on Covid recovery times. The nature of the clustering technique employed in this paper required very little precision, so this method detracts little from our process.

A.5 Splitting Data into Early and Late Surges

Armed with estimated current infectious curves for each state's data, we then defined a Covid "surge" in our context to be a period of exponential growth followed by a plateau and sudden drop off (Fig. A.3). This definition was made after a considerable amount of trial and error in the clustering that followed different, specific surge definitions. Using this plateau definition, the beginning and end of each surge for each state was identified. Since all states had a surge at the beginning of the pandemic (March 2020 for our data) and around Christmas or Thanksgiving of 2020, we considered only these surges for consistency. See tables A.1, A.2, and A.3.

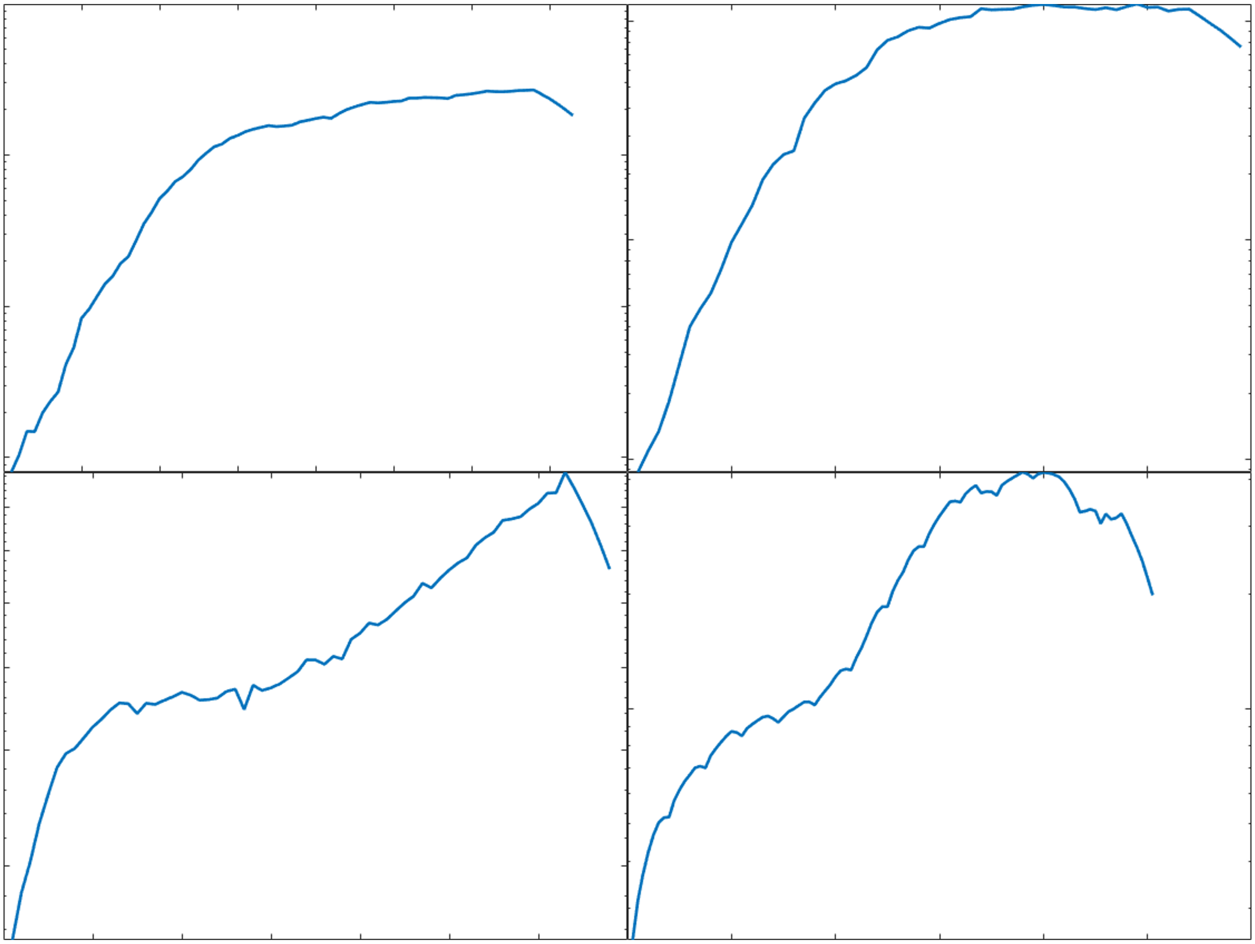


Figure A.3: Four examples of typical surges identified during the manual surge analysis. Top row: California and Georgia, early phase. Bottom row: Florida and Washington, late phase. Data shown are current infectious estimates (section A.4) shown on the log scale.

State Name	First Wave	Last Wave
AK	19-Mar-2020 to 19-Mar-2020	18-Sep-2020 to 21-Dec-2020
AL	15-Mar-2020 to 15-Mar-2020	21-Oct-2020 to 15-Feb-2021
AR	14-Mar-2020 to 14-Mar-2020	27-Oct-2020 to 30-Jan-2021
AZ	14-Mar-2020 to 14-Mar-2020	30-Sep-2020 to 05-Feb-2021
CA	04-Mar-2020 to 04-Mar-2020	31-Oct-2020 to 05-Feb-2021
CO	09-Mar-2020 to 09-Mar-2020	03-Oct-2020 to 23-Dec-2020
CT	14-Mar-2020 to 14-Mar-2020	05-Oct-2020 to 27-Dec-2020
DC	12-Mar-2020 to 12-Mar-2020	31-Oct-2020 to 05-Feb-2021
DE	17-Mar-2020 to 17-Mar-2020	31-Oct-2020 to 30-Jan-2021
FL	08-Mar-2020 to 08-Mar-2020	17-Oct-2020 to 25-Dec-2020
GA	09-Mar-2020 to 09-Mar-2020	25-Nov-2020 to 12-Feb-2021
HI	17-Mar-2020 to 17-Mar-2020	26-Dec-2020 to 09-Feb-2021
IA	11-Mar-2020 to 11-Mar-2020	24-Oct-2020 to 14-Dec-2020
ID	19-Mar-2020 to 19-Mar-2020	17-Sep-2020 to 28-Dec-2020
IL	10-Mar-2020 to 10-Mar-2020	05-Oct-2020 to 27-Dec-2020
IN	11-Mar-2020 to 11-Mar-2020	30-Sep-2020 to 29-Dec-2020

Table A.1: Estimates made for each surge time period for the first third of the data from the analysis described in section A.5

State Name	First Wave	Last Wave
KS	16-Mar-2020 to 16-Mar-2020	21-Oct-2020 to 18-Dec-2020
KY	13-Mar-2020 to 13-Mar-2020	14-Oct-2020 to 22-Dec-2020
LA	12-Mar-2020 to 12-Mar-2020	05-Nov-2020 to 24-Dec-2020
MD	12-Mar-2020 to 12-Mar-2020	30-Oct-2020 to 01-Feb-2021
ME	16-Mar-2020 to 16-Mar-2020	20-Oct-2020 to 04-Feb-2021
MI	01-Mar-2020 to 01-Mar-2020	04-Oct-2020 to 27-Dec-2020
MN	10-Mar-2020 to 10-Mar-2020	21-Oct-2020 to 22-Dec-2020
MO	18-Mar-2020 to 18-Mar-2020	29-Oct-2020 to 23-Dec-2020
MS	15-Mar-2020 to 15-Mar-2020	30-Oct-2020 to 08-Feb-2021
MT	18-Mar-2020 to 18-Mar-2020	15-Sep-2020 to 25-Dec-2020
NC	12-Mar-2020 to 12-Mar-2020	27-Sep-2020 to 17-Feb-2021
ND	19-Mar-2020 to 19-Mar-2020	14-Aug-2020 to 18-Dec-2020
NE	12-Mar-2020 to 12-Mar-2020	09-Sep-2020 to 22-Dec-2020
NH	15-Mar-2020 to 15-Mar-2020	28-Oct-2020 to 08-Feb-2021
NJ	09-Mar-2020 to 09-Mar-2020	18-Sep-2020 to 12-Feb-2021
NM	12-Mar-2020 to 12-Mar-2020	20-Sep-2020 to 22-Dec-2020

Table A.2: Estimates made for each surge time period for the second third of the data from the analysis described in section A.5

State Name	First Wave	Last Wave
NV	12-Mar-2020 to 12-Mar-2020	24-Sep-2020 to 04-Feb-2021
NY	06-Mar-2020 to 06-Mar-2020	28-Oct-2020 to 15-Feb-2021
OH	13-Mar-2020 to 13-Mar-2020	09-Oct-2020 to 29-Jan-2021
OK	16-Mar-2020 to 16-Mar-2020	01-Nov-2020 to 09-Feb-2021
OR	09-Mar-2020 to 09-Mar-2020	26-Oct-2020 to 24-Jan-2021
PA	09-Mar-2020 to 09-Mar-2020	27-Sep-2020 to 24-Jan-2021
RI	12-Mar-2020 to 12-Mar-2020	26-Sep-2020 to 26-Jan-2021
SC	12-Mar-2020 to 12-Mar-2020	25-Nov-2020 to 17-Feb-2021
SD	16-Mar-2020 to 16-Mar-2020	11-Sep-2020 to 24-Dec-2020
TN	12-Mar-2020 to 12-Mar-2020	30-Oct-2020 to 31-Jan-2021
TX	09-Mar-2020 to 09-Mar-2020	08-Nov-2020 to 14-Feb-2021
UT	15-Mar-2020 to 15-Mar-2020	20-Sep-2020 to 25-Dec-2020
VA	12-Mar-2020 to 12-Mar-2020	11-Oct-2020 to 16-Feb-2021
VT	16-Mar-2020 to 16-Mar-2020	15-Oct-2020 to 22-Dec-2020
WA	25-Feb-2020 to 25-Feb-2020	29-Oct-2020 to 08-Feb-2021
WI	13-Mar-2020 to 13-Mar-2020	05-Sep-2020 to 17-Sep-2020
WV	21-Mar-2020 to 21-Mar-2020	21-Oct-2020 to 08-Feb-2021
WY	17-Mar-2020 to 17-Mar-2020	08-Sep-2020 to 25-Dec-2020

Table A.3: Estimates made for each surge time period for the final third of the data from the analysis described in section A.5

Appendix B

Detailed Methods for COVID Modeling

B.1 Fitting an SIR Curve to Case Data

There are many statistical methods to estimate parameters from data (see e.g. [Kal60, Myu03, DJ09]). These methods often rely on assuming underlying distributions for errors in both the assumed model and the data. Reflecting our complete ignorance of the error distributions for the model and a lack of confidence in the data given, we opted for a simple approach. We derived the model parameters which produced the curve that fit the data well in a least-squares sense only. This objective is coincidentally equivalent to assuming Gaussian, time-independent error for the data, but was chosen mainly because minimizing least-squares error is a classical way of producing fits which "look" the best.

B.1.1 The Parameter Estimation Method

Suppose we have a model $Y(t, \theta)$, where t is a continuous variable and θ is a vector of parameters, and a set of time points $\{t_i\}$, $i = 1, 2, \ldots, n$, upon which data $X(t_i)$ are defined. Assume Y and X are vectors that are the same size. Fitting our model to the data means minimizing the following objective function

$$E(\theta) = \sum_{i=1}^{n} w_i ||Y(t_i, \theta) - X(t_i)||^2 \tag{B.1}$$

for some set of weights w_i, whence the best-fitting parameter vector $\hat{\theta}$ may be written as

$$\hat{\theta} = \arg\min_{\theta} E(\theta) \tag{B.2}$$

In this case, the model we desired to fit was the simple SIR model:

$$\begin{aligned}\frac{\mathbf{d}S}{\mathbf{d}t} &= -\beta SI/N_{tot} \\ \frac{\mathbf{d}I}{\mathbf{d}t} &= \beta SI/N_{tot} - \gamma I \\ \frac{\mathbf{d}R}{\mathbf{d}t} &= \gamma I\end{aligned} \tag{B.3}$$

Since data was given in terms of cumulative cases, we have $Y(t,\theta) = I(t) + R(t)$, i.e. the number of people who are currently infected plus those that have been infected. In this case $\theta = (\beta, \gamma, N_{tot}, S(0), I(0))$. The model is differentiable with respect to θ, so the minimum parameter vector may be found using gradient descent. Since the problem is non-convex, we optimized using many different initial guesses for θ in an effort to coax out the global minimum from many different local minima. In an effort to keep β realistic, we also included constrained γ to be between 0 and 1, since γ (the disease removal rate) for Covid-19 can reasonably be assumed to be in that range, and almost certainly much smaller than 1. For best visual results such as those in the clustering analysis, we let $w_i = t_i^p$ and adjusted p until the fit was to visual satisfaction.

B.1.2 Finding the Fitted β for Simulated Data

For simulated data, γ, N_{tot}, $S(0)$, and $I(0)$ were assumed to be known hyperparameters of the model, and so only $\theta = \beta$ was considered. In order to better reflect fitting that happens on real data, the continuous simulated data was sampled at intervals one day apart, and the parameters used to generate the simulated data were chosen so that the timescale of the model was approximately one day for the Covid-19 pandemic. The weights w_i were set to 1 to fit the case data equally over the entire progression of the epidemic.

B.1.3 Two-Phase SIR Model Fitting

Due to a large degree of ignorance as to what SIR initial conditions could be when fitting the model to a surge in the later stages of the pandemic, it proved necessary to allow the SIR model's susceptibility parameter to have two constant phases. The first phase, when fitting, relaxed the ignorant initial conditions to reasonable conditions for the SIR model, and the second, much longer phase captured the growth dynamics once initial conditions had been decided in the earlier phase. Specifically, we fit the model

$$\begin{aligned}\frac{\mathbf{d}S}{\mathbf{d}t} &= -\big(\beta_1 + (\beta_2-\beta_1)H(t-t^*)\big)SI/N_{tot} \\ \frac{\mathbf{d}I}{\mathbf{d}t} &= \big(\beta_1 + (\beta_2-\beta_1)H(t-t^*)\big)SI/N_{tot} - \gamma I \\ \frac{\mathbf{d}R}{\mathbf{d}t} &= \gamma I\end{aligned} \tag{B.4}$$

where $H(t)$ is the Heaviside step function, namely $H(t) = 1$ for $t > 0$ and 0 otherwise. Given nondifferentiability in the t^* parameter, a grid search was formed over many values of t^* and the minimal error solution was saved. This approach was used to generate all of the Omicron surge fits mentioned in the preceding, and invariable produced a first phase which lasted only three to five days, yielding a true single-phase SIR surge. It also better captured the behavior of the (admittedly unreliably-reported) initial pandemic data and allowed for better overall curve fit. The model fits the worst-conforming state cluster in the cluster analysis (Fig (B.1)).

B.1.4 Generating Parameter Likelihood Distributions

To generate error bars for Fig 7, we needed to define some notion of likelihood that a given set of fitted parameters is the true best-fitting set. Due to the highly nonlinear nature of the problem, many local minima exist which could reasonably approximate the data, so a likelihood was generated by randomizing the initial parameter guess and seeing which parameter minimum the algorithm arrived at. We drew our initial guess for β uniformly at random from the interval $[0.1, 0.6]$, and we drew $\log_{10} N_{tot}$ uniformly from the interval

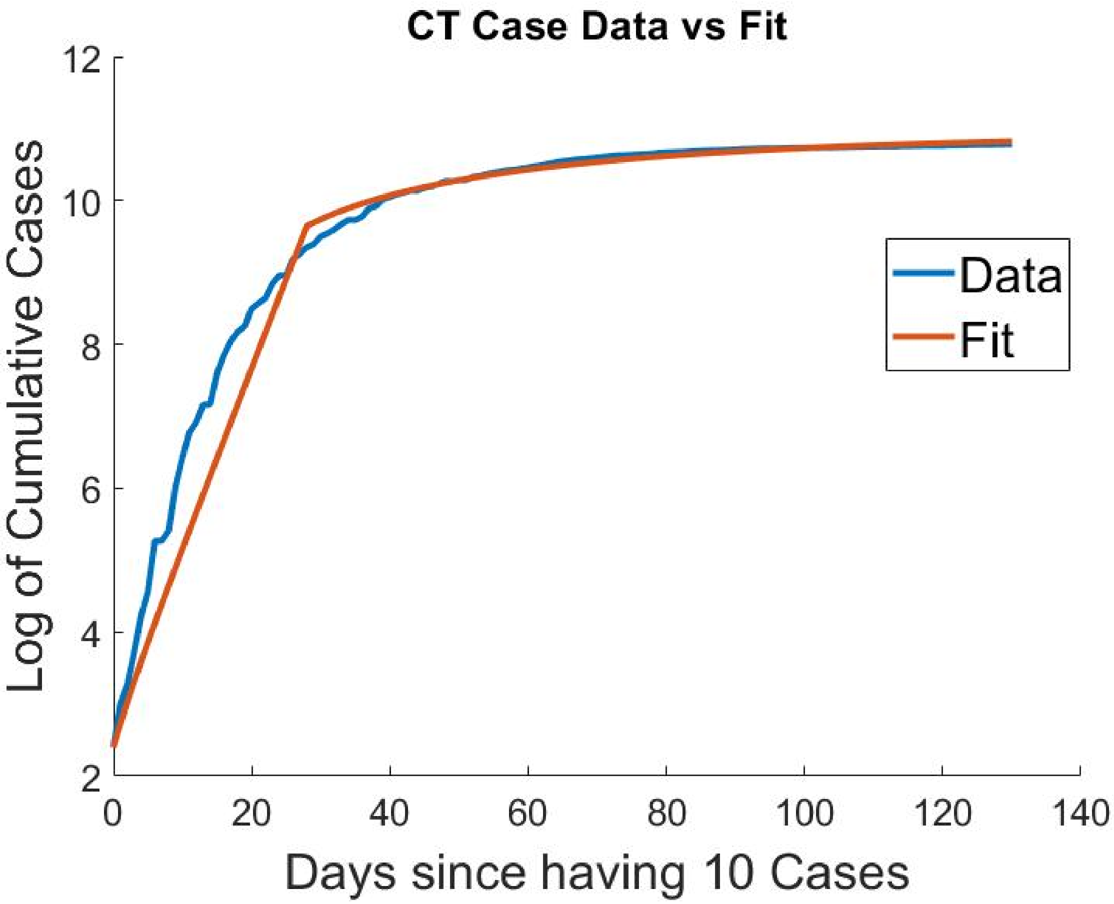


Figure B.1: Demonstration of two-phase SIR model fitting for early pandemic data. The fit to CT is vastly improved by adding a susceptibility change degree of freedom to the minimization.

[7, 8]. We then defined the likelihood of a converged-upon parameter set as e^{-BR}, where $B = 1$ and R is the mean squared residual of the model fit to the data. We found that the qualitative behavior of the error bars in Fig 7 was not sensitive to particular choices of our initial randomization or choice of B.

REFERENCES

[AAP13] Karen Alim, Gabriel Amselem, François Peaudecerf, Michael P. Brenner, and Anne Pringle. "Random network peristalsis in Physarum polycephalum organizes fluid flows across an individual." *Proceedings of the National Academy of Sciences*, **110**(33):13306–13311, 2013.

[AAP17] Karen Alim, Natalie Andrew, Anne Pringle, and Michael P Brenner. "Mechanism of signal propagation in Physarum polycephalum." *Proceedings of the National Academy of Sciences*, **114**(20):5136–5141, 2017.

[ACG20] Alex Arenas, Wesley Cota, Jesús Gómez-Gardenes, Sergio Gómez, Clara Granell, Joan T Matamalas, David Soriano-Panos, and Benjamin Steinegger. "A mathematical model for the spatiotemporal epidemic spreading of COVID19." *MedRxiv*, 2020.

[ADS20] Francesco Agostinelli, Matthias Doepke, Giuseppe Sorrenti, and Fabrizio Zilibotti. "When the Great Equalizer Shuts Down: Schools, Peers, and Parents in Pandemic Times." Working Paper 28264, National Bureau of Economic Research, December 2020.

[AHB21] Sam Abbott, Joe Hickson, Hamada S. Badr, Sebastian Funk, Pietro Monticone, Peter Ellis, J. D. Munday, Jamie Allen, Carl A. B. Pearson, Michael DeWitt, Nikos Bosse, and Sophie Meakin. "epiforecasts/EpiNow2: Beta release.", June 2021.

[AHT20] S Abbott, J Hellewell, RN Thompson, K Sherratt, HP Gibbs, NI Bosse, JD Munday, S Meakin, EL Doughty, JY Chun, YWD Chan, F Finger, P Campbell, A Endo, CAB Pearson, A Gimma, T Russell, S Flasche, AJ Kucharski, RM Eggo, and S Funk. "Estimating the time-varying reproduction number of SARS-CoV-2 using national and subnational case counts [version 2; peer review: 1 approved with reservations]." *Wellcome Open Research*, **5**(112), 2020.

[AKF16] Dai Akita, Itsuki Kunita, Mark D Fricker, Shigeru Kuroda, Katsuhiko Sato, and Toshiyuki Nakagaki. "Experimental models for Murray's law." *Journal of Physics D: Applied Physics*, **50**(2):024001, 2016.

[BBT20] Tom Britton, Frank Ball, and Pieter Trapman. "A mathematical model reveals the influence of population heterogeneity on herd immunity to SARS-CoV-2." *Science*, **369**(6505):846–849, 2020.

[BFM20] Andrea L Bertozzi, Elisa Franco, George Mohler, Martin B Short, and Daniel Sledge. "The challenges of modeling and forecasting the spread of COVID-19." *Proceedings of the National Academy of Sciences*, **117**(29):16732–16738, 2020.

[BGM07] Shweta Bansal, Bryan T Grenfell, and Lauren Ancel Meyers. "When individual behaviour matters: homogeneous and network models in epidemiology." *Journal of the Royal Society Interface*, **4**(16):879–891, 2007.

[BH13] Werner Baumgarten and Marcus JB Hauser. "Functional organization of the vascular network of Physarum polycephalum." *Physical Biology*, **10**(2):026003, 2013.

[BKA17] Felix K Bäuerle, Mirna Kramar, and Karen Alim. "Spatial mapping reveals multi-step pattern of wound healing in Physarum polycephalum." *Journal of Physics D: Applied Physics*, **50**(43):434005, 2017.

[BKS20] Devon Barrow, Nikolaos Kourentzes, Rickard Sandberg, and Jacek Niklewski. "Automatic robust estimation for exponential smoothing: Perspectives from statistics and machine learning." *Expert Systems with Applications*, **160**:113637, 2020.

[BMV12] Vincenzo Bonifaci, Kurt Mehlhorn, and Girish Varma. "Physarum can compute shortest paths." *Journal of Theoretical Biology*, **309**:121–133, 2012.

[Bro21] Ed Browne. "California, with strict mandates, has more COVID cases than Florida, Texas–here's why." *Newsweek*, 2021.

[BRP10] Shweta Bansal, Jonathan Read, Babak Pourbohloul, and Lauren Ancel Meyers. "The dynamic nature of contact networks in infectious disease epidemiology." *Journal of biological dynamics*, **4**(5):478–489, 2010.

[CFF13] Anne Cori, Neil M. Ferguson, Christophe Fraser, and Simon Cauchemez. "A New Framework and Software to Estimate Time-Varying Reproduction Numbers During Epidemics." *American Journal of Epidemiology*, **178**(9):1505–1512, 09 2013.

[Cor15] Stefano Coraluppi. "Fundamentals and Advances in Multiple-Hypothesis Tracking." In *NATO STO IST-134 Lecture Series on Advanced Algorithms for Effectively Fusing Hard and Soft Information*. NATO Collaboration and Support Office, 2015.

[Cur] Christopher Curley. "Why Do California and Florida Have Similar Covid-19 Case Rates? The Answer Is Complicated." *HealthLine*.

[DCO18] John Dahl, Gabriel Rodrigues de Campos, Claes Olsson, and Jonas Fredriksson. "Collision avoidance: A literature review on threat-assessment techniques." *IEEE Transactions on Intelligent Vehicles*, **4**(1):101–113, 2018.

[DH00] Odo Diekmann and Johan Andre Peter Heesterbeek. *Mathematical epidemiology of infectious diseases: model building, analysis and interpretation*, volume 5. John Wiley & Sons, 2000.

[DHM90] Odo Diekmann, Johan Andre Peter Heesterbeek, and Johan AJ Metz. "On the definition and the computation of the basic reproduction ratio R 0 in models for infectious diseases in heterogeneous populations." *Journal of mathematical biology*, **28**(4):365–382, 1990.

[DJ09] Arnaud Doucet and Adam M Johansen. "A tutorial on particle filtering and smoothing: Fifteen years later." *Handbook of nonlinear filtering*, **12**(656-704):3, 2009.

[DK01] Iain S Duff and Jacko Koster. "On algorithms for permuting large entries to the diagonal of a sparse matrix." *SIAM Journal on Matrix Analysis and Applications*, **22**(4):973–996, 2001.

[DLB10] Audrey Dussutour, Tanya Latty, Madeleine Beekman, and Stephen J Simpson. "Amoeboid organism solves complex nutritional challenges." *Proceedings of the National Academy of Sciences*, **107**(10):4607–4611, 2010.

[DW02] Pauline Van den Driessche and James Watmough. "Reproduction numbers and sub-threshold endemic equilibria for compartmental models of disease transmission." *Mathematical biosciences*, **180**(1-2):29–48, 2002.

[EK02] Ken TD Eames and Matt J Keeling. "Modeling dynamic and network heterogeneities in the spread of sexually transmitted diseases." *Proceedings of the national academy of sciences*, **99**(20):13330–13335, 2002.

[Epi] "Epiforecasts." `https://epiforecasts.io/covid/posts/subnational/united-states/california/`.

[Fet22] Matt Fetterman. "Anaskel writtein in Matlab.", 2022.

[FKW22] Philipp Fleig, Mirna Kramar, Michael Wilczek, and Karen Alim. "Emergence of behaviour in a self-organized living matter network." *eLife*, **11**:e62863, jan 2022.

[FWD17] Aden Forrow, Francis G. Woodhouse, and Jörn Dunkel. "Mode Selection in Compressible Active Flow Networks." *Physical Review Letters*, **119**:028102, Jul 2017.

[FWD18] Aden Forrow, Francis G Woodhouse, and Jörn Dunkel. "Functional control of network dynamics using designed Laplacian spectra." *Physical Review X*, **8**(4):041043, 2018.

[GC78] A Grebecki and M Cieślawska. "Plasmodium of Physarum polycephalum as a synchronous contractile system." *Cytobiologie*, **17**(2):335–342, 1978.

[Het00a] Herbert W. Hethcote. "The Mathematics of Infectious Diseases." *SIAM Review*, **42**(4):599–653, 2000.

[Het00b] Herbert W Hethcote. "The mathematics of infectious diseases." *SIAM review*, **42**(4):599–653, 2000.

[HJP21] Cecilie Bo Hansen, Ida Jarlhelt, Laura Pérez-Alós, Lone Hummelshøj Landsy, Mette Loftager, Anne Rosbjerg, Charlotte Helgstrand, Jais Rose Bjelke, Thomas Egebjerg, Joseph G. Jardine, Charlotte Sværke Jørgensen, Kasper Iversen, Rafael Bayarri-Olmos, Peter Garred, and Mikkel-Ole Skjoedt. "SARS-CoV-2 Antibody Responses Are Correlated to Disease Severity in Covid-19 Convalescent Individuals." *The Journal of Immunology*, **206**(1):109–117, 2021.

[IM20a] IHME COVID-19 health service utilization forecasting team and Christopher JL Murray. "Forecasting Covid-19 impact on hospital bed-days, ICU-days, ventilator-days and deaths by US state in the next 4 months." *medRxiv*, 2020.

[IM20b] IHME COVID-19 health service utilization forecasting team and Christopher JL Murray. "Forecasting the impact of the first wave of the Covid-19 pandemic on hospital demand and deaths for the USA and European Economic Area countries." *medRxiv*, 2020.

[IR21] Nicholas J. Irons and Adrian E. Raftery. "Estimating SARS-CoV-2 infections from deaths, confirmed cases, tests, and random surveys." *Proceedings of the National Academy of Sciences*, **118**(31), 2021.

[KA21] Mirna Kramar and Karen Alim. "Encoding memory in tube diameter hierarchy of living flow network." *Proceedings of the National Academy of Sciences*, **118**(10):e2007815118, 2021.

[Kal60] Rudolph Emil Kalman. "A New Approach to Linear Filtering and Prediction Problems." *Transactions of the ASME–Journal of Basic Engineering*, **82**(Series D):35–45, 1960.

[Kam81] Nobur Kamiya. "Physical and chemical basis of cytoplasmic streaming." *Annual Review of Plant Physiology*, **32**(1):205–236, 1981.

[KAY88] N Kamiya, RD Allen, and Y Yoshimoto. "Dynamic organization of Physarum plasmodium." *Cell Motility and the Cytoskeleton*, **10**(1-2):107–116, 1988.

[KE05] Matt J Keeling and Ken TD Eames. "Networks and epidemic models." *Journal of the Royal Society Interface*, **2**(4):295–307, 2005.

[Kee05] Matt Keeling. "The implications of network structure for epidemic dynamics." *Theoretical population biology*, **67**(1):1–8, 2005.

[KNS21] Norihito Kaku, Fumitaka Nishimura, Yui Shigeishi, Rina Tachiki, Hironori Sakai, Daisuke Sasaki, Kenji Ota, Kei Sakamoto, Kosuke Kosai, Hiroo Hasegawa, Koichi Izumikawa, Koya Ariyoshi, Hiroshi Mukae, Jiro Yasuda, Kouichi Morita, Shigeru Kohno, and Katsunori Yanagihara. "Performance of anti-SARS-CoV-2 antibody testing in asymptomatic or mild Covid-19 patients: A retrospective study in outbreak on a cruise ship." *PLOS ONE*, **16**(9):1–12, 09 2021.

[KRM97] MJ Keeling, DA Rand, and AJ Morris. "Correlation models for childhood epidemics." *Proceedings of the Royal Society of London. Series B: Biological Sciences*, **264**(1385):1149–1156, 1997.

[LB11] Tanya Latty and Madeleine Beekman. "Speed–accuracy trade-offs during foraging decisions in the acellular slime mould Physarum polycephalum." *Proceedings of the Royal Society B: Biological Sciences*, **278**(1705):539–545, 2011.

[Lew05] Roger R Lew. "Mass flow and pressure-driven hyphal extension in Neurospora crassa." *Microbiology*, **151**(8):2685–2692, 2005.

[Li20] Tianyi Li. "Simulating the spread of epidemics in China on multi-layer transportation networks: Beyond COVID-19 in Wuhan." *EPL (Europhysics Letters)*, **130**(4):48002, jun 2020.

[LKK21] H. Lau, T. Khosrawipour, P. Kocbach, H. Ichii, J. Bania, and V. Khosrawipour. "Evaluating the massive underreporting and undertesting of COVID-19 cases in multiple global epicenters." *Pulmonology*, **27**(2):110–115, 2021.

[LM] Rong-Gong Lin II and Luke Money. "California doing much better with Delta variant than Florida, Texas. Here's why." *Los Angeles Times*.

[LZG15] Owen L Lewis, Shun Zhang, Robert D Guy, and Juan C Del Alamo. "Coordination of contractility, adhesion and flow in migrating Physarum amoebae." *Journal of The Royal Society Interface*, **12**(106):20141359, 2015.

[MAN17] Bernd Meyer, Cedrick Ansorge, and Toshiyuki Nakagaki. "The role of noise in self-organized decision making by the true slime mold Physarum polycephalum." *PloS one*, **12**(3):e0172933, 2017.

[MHS20] Christina Maxouris, Jason Hanna, Kristina Sgueglia, and Steve Almasy. "As of Wednesday, every state will be somewhere along the road toward a full reopening." *CNN*, May 2020.

[MM08] Patrick N McGraw and Michael Menzinger. "Laplacian spectra as a diagnostic tool for network structure and dynamics." *Physical Review E*, **77**(3):031102, 2008.

[MMZ20] Raul Macias Gil, Jasmine R Marcelin, Brenda Zuniga-Blanco, Carina Marquez, Trini Mathew, and Damani A Piggott. "COVID-19 pandemic: disparate health impact on the Hispanic/Latinx population in the United States." *The Journal of infectious diseases*, **222**(10):1592–1595, 2020.

[MPV02] Yamir Moreno, Romualdo Pastor-Satorras, and Alessandro Vespignani. "Epidemic outbreaks in complex heterogeneous networks." *The European Physical Journal B-Condensed Matter and Complex Systems*, **26**(4):521–529, 2002.

[MTN08] Kenji Matsumoto, Seiji Takagi, and Toshiyuki Nakagaki. "Locomotive mechanism of Physarum plasmodia based on spatiotemporal analysis of protoplasmic streaming." *Biophysical journal*, **94**(7):2492–2504, 2008.

[MVZ21] Seyed M Moghadas, Thomas N Vilches, Kevin Zhang, Chad R Wells, Affan Shoukat, Burton H Singer, Lauren Ancel Meyers, Kathleen M Neuzil, Joanne M Langley, Meagan C Fitzpatrick, et al. "The impact of vaccination on coronavirus disease 2019 (COVID-19) outbreaks in the United States." *Clinical Infectious Diseases*, **73**(12):2257–2264, 2021.

[Myu03] In Jae Myung. "Tutorial on maximum likelihood estimation." *Journal of Mathematical Psychology*, **47**(1):90–100, 2003.

[ND21] Jungsik Noh and Gaudenz Danuser. "Estimation of the fraction of Covid-19 infected people in U.S. states and countries worldwide." *PLOS ONE*, **16**(2):1–10, 02 2021.

[NG08] Toshiyuki Nakagaki and Robert D Guy. "Intelligent behaviors of amoeboid movement based on complex dynamics of soft matter." *Soft Matter*, **4**(1):57–67, 2008.

[NYT00] Toshiyuki Nakagaki, Hiroyasu Yamada, and Ágota Tóth. "Maze-solving by an amoeboid organism." *Nature*, **407**(6803):470–470, 2000.

[NYU00] Toshiyuki Nakagaki, Hiroyasu Yamada, and Tetsuo Ueda. "Interaction between cell shape and contraction pattern in the Physarum plasmodium." *Biophysical chemistry*, **84**(3):195–204, 2000.

[OYS21] Kenji Ota, Katsunori Yanagihara, Daisuke Sasaki, Norihito Kaku, Naoki Uno, Kei Sakamoto, Kosuke Kosai, Taiga Miyazaki, Hiroo Hasegawa, Ayumi Fujita, Masato Tashiro, Takeshi Tanaka, Koichi Izumikawa, Koya Ariyoshi, Hiroshi Mukae, Jiro Yasuda, Kouichi Morita, and Shigeru Kohno. "Detection of SARS-CoV-2 using qRT-PCR in saliva obtained from asymptomatic or mild Covid-19 patients, comparative analysis with matched nasopharyngeal samples." *PLOS ONE*, **16**(6):1–9, 06 2021.

[RDT15] Jean-Paul Rieu, Hélene Delanoë-Ayari, Seiji Takagi, Yoshimi Tanaka, and Toshiyuki Nakagaki. "Periodic traction in migrating large amoeba of Physarum polycephalum." *Journal of The Royal Society Interface*, **12**(106):20150099, 2015.

[Rey20] Maritza Vasquez Reyes. "The disproportional impact of COVID-19 on African Americans." *Health and human rights*, **22**(2):299, 2020.

[RMM16] Chris R Reid, Hannelore MacDonald, Richard P Mann, James AR Marshall, Tanya Latty, and Simon Garnier. "Decision-making without a brain: How an amoeboid organism solves the two-armed bandit." *Journal of The Royal Society Interface*, **13**(119):20160030, 2016.

[RS19] Marcus Roper and Agnese Seminara. "Mycofluidics: the fluid mechanics of fungal adaptation." *Annual Review of Fluid Mechanics*, **51**:511–538, 2019.

[RTU15] Beatrice Rodiek, Seiji Takagi, Tetsuo Ueda, Marcus Hauser, et al. "Patterns of cell thickness oscillations during directional migration of Physarum polycephalum." *European Biophysics Journal*, **44**(5):349–358, 2015.

[Sok58] Robert R Sokal. "A statistical method for evaluating systematic relationships." *Univ. Kansas, Sci. Bull.*, **38**:1409–1438, 1958.

[SPW20] Brendan Saloner, Kalind Parish, Julie A Ward, Grace DiLaura, and Sharon Dolovich. "COVID-19 cases and deaths in federal and state prisons." *Jama*, **324**(6):602–603, 2020.

[SSR19] Stefanie S Schmieder, Claire E Stanley, Andrzej Rzepiela, Dirk van Swaay, Jerica Sabotič, Simon F Nørrelykke, Andrew J deMello, Markus Aebi, and Markus Künzler. "Bidirectional propagation of signals and nutrients in fungal networks via specialized hyphae." *Current Biology*, **29**(2):217–228, 2019.

[SSV13] Faryad Darabi Sahneh, Caterina Scoglio, and Piet Van Mieghem. "Generalized epidemic mean-field model for spreading processes over multilayer complex networks." *IEEE/ACM Transactions on Networking*, **21**(5):1609–1620, 2013.

[The] "The COVID Tracking Project." `https://covidtracking.com/`.

[TL20] Juliana Tolles and ThaiBinh Luong. "Modeling epidemics with compartmental models." *Jama*, **323**(24):2515–2516, 2020.

[TPS18] Ayushi Trivedi, Navya Pant, Pinal Shah, Simran Sonik, and Supriya Agrawal. "Speech to text and text to speech recognition systems-A review." *IOSR J. Comput. Eng*, **20**(2):36–43, 2018.

[TSv19] R.N. Thompson, J.E. Stockwin, R.D. van Gaalen, J.A. Polonsky, Z.N. Kamvar, P.A. Demarsh, E. Dahlqwist, S. Li, E. Miguel, T. Jombart, J. Lessler, S. Cauchemez, and A. Cori. "Improved inference of time-varying reproduction numbers during infectious disease outbreaks." *Epidemics*, **29**:100356, 2019.

[TTS10] Atsushi Tero, Seiji Takagi, Tetsu Saigusa, Kentaro Ito, Dan P Bebber, Mark D Fricker, Kenji Yumiki, Ryo Kobayashi, and Toshiyuki Nakagaki. "Rules for biologically inspired adaptive network design." *Science*, **327**(5964):439–442, 2010.

[UMA86] Tetsuo Ueda, Kenji Matsumoto, Tatsuo Akitaya, and Yonosuke Kobatake. "Spatial and temporal organization of intracellular adenine nucleotides and cyclic nucleotides in relation to rhythmic motility in Physarum plasmodium." *Experimental cell research*, **162**(2):486–494, 1986.

[VM07] Erik Volz and Lauren Ancel Meyers. "Susceptible–infected–recovered epidemics in dynamic contact networks." *Proceedings of the Royal Society B: Biological Sciences*, **274**(1628):2925–2934, 2007.

[WFF16] Francis G Woodhouse, Aden Forrow, Joanna B Fawcett, and Jörn Dunkel. "Stochastic cycle selection in active flow networks." *Proceedings of the National Academy of Sciences*, **113**(29):8200–8205, 2016.

[Wil22] Ryan Wilkinson. "'Flow modes provide a quantification of Physarum network peristalsis Code Supplement." https://data.mendeley.com/datasets/2ntjd49s7z/1, 2022.

[Woo] John Woolfolk. "California's Covid-19 case rate now twice Florida's." *Tampa Bay Times*.

[Woo21] John Woodfolk. "COVID: What do California, Texas, New York and Florida have in common? Stunningly low infection rates." *The Mercury News*, 2021.

[WTT11] Shin Watanabe, Atsushi Tero, Atsuko Takamatsu, and Toshiyuki Nakagaki. "Traffic optimization in railroad networks using an algorithm mimicking an amoeba-like organism, Physarum plasmodium." *Biosystems*, **105**(3):225–232, 2011.

[Yan08] Ping Yan. "Distribution theory, stochastic processes and infectious disease modelling." In *Mathematical epidemiology*, pp. 229–293. Springer, 2008.

[YK84] Yasuaki Yoshimoto and Noburo Kamiya. "ATP-and calcium-controlled contraction in a saponin model of Physarum polycephalum." *Cell Structure and Function*, **9**(2):135–141, 1984.

[YLI20] Karen Yourish, KK Rebecca Lai, Danielle Ivory, and Mitch Smith. “One-third of all US coronavirus deaths are nursing home residents or workers.” *New York Times*, **5**(9), 2020.

[ZGL17] Shun Zhang, Robert D Guy, Juan C Lasheras, and Juan C Del Álamo. “Self-organized mechano-chemical dynamics in amoeboid locomotion of Physarum fragments.” *Journal of physics D: Applied physics*, **50**(20):204004, 2017.