UNIVERSITY OF CALIFORNIA,
IRVINE

The Effects of Geography on Social Structures

DISSERTATION

submitted in partial satisfaction of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

in Sociology

by

Loring J. Thomas

Dissertation Committee:
Chancellor's Professor Carter T. Butts, Chair
Professor John R. Hipp
Professor Emerita Katherine Faust

2023

DEDICATION

To my parents, who supported and encouraged me.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Page

LIST OF ALGORITHMS

ACKNOWLEDGMENTS

I would like to thank my faculty advisor Carter T. Butts for his help and sagely advice throughout graduate school. Chapter 2 of this thesis includes a reprint of materials published in PNAS [Thomas et al., 2020, 2022a], used with permission from PNAS. The coauthors for these publications are: Carter T. Butts, John R. Hipp, Zack Almquist, Peng Huang, Fan Yin, Iris Xiaoshuang Luo, and Junlan Xu. Chapter 3 of my dissertation is coauthored with Peng Huang, Iris Xiaoshuang Luo, John R. Hipp, and Carter T. Butts, and has been conditionally accepted at *Sociological Methodology*. Chapter 4 is unpublished research in collaboration with Carter T. Butts. I would like to thank my friends, who supported me on my graduate school journey, and my wonderful partner Thoa Khuu for her support and advice. I would like to thank the Proceedings of the National Academy of Science for their policies allowing the inclusion of the two papers making up Chapter 2 in my dissertation. This research was funded by NSF Awards IIS-1939237, SES-1826589 and a University of California, Irvine Council on Research, Computing and Libraries grant (all awarded to Carter T. Butts), and NIH Award P2C HD042828 to the Center for Studies in Demography and Ecology for Zack W. Almquist for chapter 2, award SES-1826589 for chapter 3, and NSF award SES-1826589 and NIH award 1R01GM144964-01 for chapter 4.

VITA

Loring J. Thomas

EDUCATION

Doctor of Philosophy in Sociology **2023**
University of California, Irvine *Irvine, California*

Master of Arts in Social Science **2020**
University of California, Irvine *Irvine, California*

Bachelor of Arts in Computer Science and Sociology **2017**
Pitzer College *Claremont, California*

RESEARCH EXPERIENCE

Graduate Research Assistant **2018–2023**
University of California, Irvine *Irvine, California*

TEACHING EXPERIENCE

Teaching Assistant **2017–2018**
University of California, Irvine *Irvine, California*

REFEREED JOURNAL PUBLICATIONS

Spatial heterogeneity can lead to substantial local variations in COVID-19 timing and severity **2020**
Loring J. Thomas, Peng Huang, Fan Yin, Iris Xiaoshuang Luo, Zack W. Almquist, John R. Hipp, Carter T. Butts, PNAS

Geographical patterns of social cohesion drive disparities in early COVID infection hazard **2022**
Loring J. Thomas, Peng Huang, Fan Yin, Junlan Xu, Zack W. Almquist, John R. Hipp, Carter T. Butts, PNAS

ABSTRACT OF THE DISSERTATION

The Effects of Geography on Social Structures

By

Loring J. Thomas

Doctor of Philosophy in Sociology

University of California, Irvine, 2023

Chancellor's Professor Carter T. Butts, Chair

Physical space and geography have a significant effect on social structures. This dissertation discusses social networks, which describe the relational structure of a group of entities. In addition, I examine nested spatial data hierarchies. I lead the following research, along with collaborators, to examine three different ways in which social networks and other social systems that are embedded in the world are affected by geography and distance.

Chapter 2 of the dissertation examines the mechanisms through which geographic variability affects the spread of COVID-19. We simulate large scale social contact networks for the city of San Francisco and simulate the spread of COVID-19 across these networks, using mortality and case data from early in the pandemic. We find that local social cohesion is a hidden risk factor for the spread of COVID-19 in urban environments. For wild-type COVID-19, individual degree (i.e. the number of social contacts one has), was not as important in explaining the spread of the disease as local cohesion (i.e. how connected your neighbors are to each other). This research also highlights the spatial heterogeneities in COVID spread, where many parts of the city did not experience significant COVID related health outcomes until well after the disease began to spread as a result of the clustered nature of the social networks underlying disease spread.

In Chapter 3 of the dissertation, we implement and validate a marginal preserving imputation algorithm to impute three-way crosstab data that is embedded in spatial hierarchies. Spatial data hierarchies, such as the U.S. Census or Google S2 Geometry systems, contain nested sets of areal units. Each level of the spatial hierarchy is composed of aggregated areal units at a lower level of geography. For example, census tracts are composed on a set of census blocks. These data are valuable, as they often are collected at large scales and a variety of degrees of spatial resolution. Despite that, for data collection issues or other reasons, some levels of geography may be missing or unavailable at the three-way level. Given a set of known two way marginals at the target level of imputation, and a full set of three-way data at a higher level of geography, we

implement a marginal preserving imputation algorithm that maintains all known two-way marginals at the target level, while preserving higher order correlations between cells in the three-way array with data from the higher level of geography. This imputation algorithm uses MCMC and simulated annealing to optimize the state of a three-way array to the distribution observed at the higher level of geography. We impute the three-way arrays for the distribution of population by race, ethnicity, and gender in census tracts across the US, and census blocks across California. We validate the imputed data quality by comparing imputed three-way arrays for census tracts with their observed counterparts. A case study shows that downstream analyses using imputed arrays is not likely to significantly impact conclusions.

Chapter 4 of the dissertation develops an extension of the ERGM framework. Exponential Family Random graph Models (ERGM) model edge probabilities in a network as a function of a set of sufficient statistics. When modeling dynamic social networks, most researchers hold the vertex set, or the people/organizations participating in the social network, constant. We integrate Generalized Location Systems (GLS) into the ERGM framework, which allows for the separable parameterization of vertex and edge processes. This new model framework allows us to model endogenous patterns in network participation in addition to the endogenous edge processes that the ERGM framework can currently model. We develop a quorum of model terms to describe the movement of nodes in and out of the social network. These terms allows for the parameterization of effects for covariates, endogenous participation, and other network effects into the vertex dynamics model. Using the Freeman Windsurfer data, we parameterize and fit several models. We use standard model adequacy techniques (one-step model prediction) to validate the quality of these models. This modeling framework provides new opportunities to model complex network systems in which the process of nodes entering and exiting the network is of interest. These networks include voluntary organizations, social movements, and endogenous groups.

Chapter 1

Introduction

The distribution of people across space is heterogenous, and within the United States are often segregated by race, income, and other demographic attributes. Significant clusters of people exist near cities, but the distribution of populations across these spaces is not consistent. Variations in population distribution by race, ethnicity, age, income and other demographic covariates exist due to the segregated nature of many American cities [Massey et al., 2009]. Massey notes that over the last century, the scale of segregation has decreased, with segregation occurring at finer and finer levels of geography, while also changing to be more centered on income segregation that race. The study of how geography and the uneven distribution of people across space affect social outcomes thus remains an important problem to examine, especially when researchers are able to integrate fine geographic detail into their work.

While work on segregation has highlighted the importance of a focus on geography, many other subfields also have benefitted from this spatial lens. Studies of patterns in marriage, city size, formation of social ties, epidemiology, diffusion of ideas, social network structure, and regional identification have all shown that geography and the spatial structure of the United States are important factors in understanding social phenomena [Bossard, 1938, Zipf, 2016, Butts and Acton, 2011, Thomas et al., 2020, Boessen et al., 2017, Almquist and Butts, 2015, Festinger et al., 1950]. This body of work highlights the importance of understanding how physical space can structure social relations, but also in how social outcomes can be suppressed through spatial distance.

Given that geography has been identified as an important predictor for a variety of social outcomes, this

dissertation examines several different ways in which space and geography affect social outcomes. In the first chapter, we examine how integrating insights from spatial network structures can capture the effects of spatial heterogeneity on disease spread in the context of COVID-19. We highlight the value of using spatial social networks to examine disease spread, and show that the cohesion of one's local network can affect infection hazard at the beginning of the pandemic. The second chapter concerns the distribution of populations across space. We implement and test a novel three-way imputation technique for data embedded within spatial hierarchies. Our method allows for the integration of higher level spatial information about population structure into the imputation of data at more local levels. Finally, the third chapter introduces a novel modeling framework for social networks that are in demographic exchange with their environments (i.e. the vertex set of the network changes over time). This chapter examines the ways in which networks that are present in a specific location may change over time as people enter and exit then network. The main contributions of this thesis are in solidfying the effects of geography on these social systems, but also in providing tools with which to model spatial processes related to disease, population distributions, and social network structure.

Chapter 2

The Effects of Geography and Distance on COVID-19 Spread in Social Networks

2.1 Introduction

Since its emergence at the end of 2019, the SARS-CoV-2 virus has spread rapidly to all portions of globe, infecting over 750 million people as of Spring 2023 [World Health Organization, 2023]. The disease caused by this virus, denoted COVID-19, generally manifests as a respiratory illness that is spread primarily via airborne droplets. While most cases of COVID-19 are non-fatal, a significant fraction of those infected require extensive supportive care, and the mortality rate is substantially higher than more common infectious diseases such as seasonal influenza [Onder et al., 2020]. Even for survivors, infection can lead to long-term damage to the lungs and other organs, leading to long convalescence times and enhance risks of secondary complications [Jiang et al., 2020, Geng et al., 2020]. By early March of 2020, COVID-19 outbreaks had appeared on almost every continent, including significant clusters within many cities [World Health Organization, 2020a]. Prior to the development and public release of a vaccine in late 2020, public health measures to counteract the pandemic in developed nations have focused on social distancing measures that seek to slow diffusion sufficiently to avoid catastrophic failure of the healthcare delivery system. Both the planning and public

acceptance of such measures have been highly dependent upon the use of epidemiological models to probe the potential impact of distancing interventions, and to anticipate when such measures may be loosened with an acceptable level of public risk. As such, the assumptions and behavior of COVID-19 diffusion models is of significant concern.

When the pandemic first came to the attention of the epidemiological community, dominant approaches to COVID-19 modeling [Jackson et al., 2020, Zhang et al., 2020, Pujari and Shekatkar, 2020] are based on compartment models (often called *SIR* models, after the conventional division of the population into *susceptible*, *infected*, and *recovered* groups in the most basic implementations) that implicitly treat individuals within a population as geographically well-mixed. While some such models include differential contact by demographic groups (e.g., age), and may treat states, counties, or occasionally cities as distinct units (e.g. work by [Brockmann and Helbing, 2013]), those models presently in wide use do not incorporate spatial heterogeneity at local scales (e.g., within cities). Past work, however, has shown evidence of substantial heterogeneity in social relationships at regional, urban, and sub-urban scales [Spiro et al., 2016, Smith et al., 2015, Wang et al., 2018], with these variations in social network structure impacting outcomes as diverse as regional identification Almquist and Butts [2015], disease spread [Riley, 2007], crime rates [Hipp et al., 2013], neighborhood identification and development [Sampson and Sharkey, 2008, Wang et al., 2018]. If individuals are not socially "well-mixed" at local scales, then it is plausible that diffusion of SARS-CoV-2 via interpersonal contacts will likewise depart from the uniform mixing characteristic of the SIR models. Indeed, at least one computational study [Almquist and Butts, 2012] using a fairly "generic" (non-COVID) diffusion process on realistic urban networks has showed considerable non-uniformity in diffusion times, suggesting that such effects could hypothetically be present. Variations across local regions on the pandemic timing, severity, and the hospital load could have huge impacts on the social outcomes of different population groups (e.g. racial/ethnic groups) in the pandemic, given the heterogeneity of their spatial distribution in urban and sub-urban areas [Massey and Denton, 1993, Massey and Tannen, 2018]. However, it could also be hypothesized that such effects would be small perturbations to the broader infection curve captured by conventional compartment models, with little practical importance. In the opening months of the pandemic, this remained an open question, although this research and observations about the trajectory of the pandemic indicate that there is significant heterogeneity in COVID trajectories and impacts across space and social groups.

In this chapter, we examine the potential impact of local spatial heterogeneity on COVID-19, modeling the diffusion of SARS-CoV-2 in populations whose contacts are based on spatially plausible network structures

and assessing potential network mechansims that could contribute to heterogeneous outcomes. We focus here on the urban context, examining nineteen different cities in the United States at a broad level, as well as examining the city of San Francisco far more in depth. We simulate the population of each city in detail (i.e., at the individual level), simulating hypothetical outbreaks on the contact network in each city in the absence of measures such as social distancing. Despite allowing the population to be well-mixed in all other respects (i.e., not imposing mixing constraints based on demographic or other characteristics), we find that spatial heterogeneity alone is sufficient to induce substantial departures from spatially homogeneous SIR behavior. Among the phenomena observed are "long lag" outbreaks that appear in previously unharmed communities after the aggregate infection wave has largely subsided; frequently low correlations between infection timing in spatially adjacent communities; and distinct sub-patterns of outbreaks found in some urban areas that are uncorrelated with the broader infection pattern. Gaps between infection peaks at the intra-urban level can be large, e.g. on the order of weeks or months in extreme cases, even for communities that are within kilometers of each other. Such heterogeneity is potentially consequential for the management of healthcare delivery services: as we show using a simple "catchment" model of hospital demand, local variations in infection timing can easily overload hospitals in some areas, generating "hospital deserts" Verhagen et al. [2020], while leaving others relatively empty (absent active reallocation of patients). Likewise, we show that individuals' social exposures to others who are morbid or deceased vary greatly over the course of the pandemic, potentially leading to differences in risk assessment and bereavement burden for persons residing in different locations. Differences in outbreak timing and severity may exacerbate health disparities (since e.g., surge capacity varies by community) and may even affect perception of and support for prophylactic behaviors among the population at large, with those in so-far untouched communities falsely assuming that the pandemic threat is either past or was exaggerated to begin with, or attributing natural variation in disease timing to the impact of health interventions. Given the surge in prostests against non-pharmaceutical interventions (NPIs) across the first year of the pandemic, spatial exposure to different COVID-19 infection and death rates may have had a contribution.

While the main focus of this first chapter is on understanding the effect of heterogeneous population distributions on COVID-19 spread, much of this work was completed very early in the pandemic. As the pandemic continues, many of the findings presented were supported by the observed trajectory of the pandemic. For example, the long tailed nature of the pandemic, while not obvious at the time, has been supported, as COVID-19 infections are still a salient matter years after the initial outbreak. An important feature of the pandemic that was highlighted was the heterogeneous impact of COVID-19 across race and ethnicity [CDC, 2021, Magesh et al., 2021]. In addition to the large scale simulations that we run, highlighting the effects of

population distributions on the heterogeneous trajectory of the pandemic, we also conduct an investigation into the potential mechanistic underpinnings of this finding, as the mechanistic connections between contact network structure and infection hazard are not fully understood. Here, we show that small differences in local *social cohesion* can result in large disparities in infection rates by race and ethnicity as observed in the U.S. While long-term outcomes are important, we specifically aim to understand how the disparities in infection by race and ethnicity arise *early* in the pandemic. In the initial phase of an emerging pandemic, risks are unclear, non-pharmaceutical interventions (e.g., masking, distancing) are not yet implemented, and behavioral changes are rarely widespread; yet, it is precisely at this point that the virus has the greatest opportunity to penetrate the population, with the capacity to provide particular harms to vulnerable communities.

We examine the structural precedents of the heterogeneity in infection rates by race and ethnicity by proposing a mechanism through which the structure of the social network affects the patterns of COVID-19 spread. While in general, the epidemiological literature has shown that population density increases the rate of disease spread [Kadi and Khelfaoui, 2020, Rashed et al., 2020], it does not provide a mechanistic interpretation for this phenomenon. However, previous research on spatial network models has highlighted the way in which density can drive tie creation and resulting cohesive subgroup formation [Butts et al., 2012a]. Our models provide a specific mechanism for how population density and household size distributions may result in increased disease spread: population distribution influences the creation of locally cohesive regions within the contact network, and these regions are *exceptionally permeable* to SARS-CoV-2.

We note at the outset that the models used here are intended to probe the hypothetical impact of spatial heterogeneity on COVID-19 diffusion within particular scenarios, rather than to produce high-accuracy predictions or forecasts. While several years after the beginning of the pandemic, many of the qualitative predictions produced by the simulations we descibe have been supported, the trajectories generated in this work are from baseline models of social network generation and COVID-19 spread. For high-accuracy predictions and forecasts, it is desirable to incorporate many additional features that are here simplified to facilitate insight into the phenomenon of central interest. In particular, we do not incorporate either many demographic effects or social distancing [Dowd et al., 2020, Block et al., 2020], allowing us consider a setting that is as well-mixed as possible (and hence as close as possible to an idealized SIR model) with the exception of spatial heterogeneity. As we show, even this basic scenario is sufficient to produce large deviations from the SIR model. Despite the simplicity of our models, we do note that the approach employed here could be integrated with other factors and calibrated to produce models intended for forecasting or similar applications.

2.2 Methods

COVID-19 is typically transmitted via direct contact with infected individuals, with the greatest risk occurring when an uninfected person is within approximately six feet of an infected person for an extended period of time. Such interactions can be modeled as events within a social network, where individuals are tied to those whom they have a high hazard of intensive interaction. In prior work, this approach has been successfully employed for modeling infectious diseases ranging from HIV [Morris, 2004] and influenza [Viboud et al., 2006] to Zika [Li et al., 2019]. To model networks of potential contacts at scale, we employ spatial network models Butts and Acton [2011], which are both computationally tractable and able to capture the effects of geography and population heterogeneity on network structure [Butts et al., 2012b]. Such models have been successfully used to capture social phenomena ranging from neighborhood-level variation crime rates [Hipp et al., 2013] and regional identification [Almquist and Butts, 2015] to the flow of information among homeless persons [Almquist, 2020].

The spatial network models used here allow for complex social dependence through a kernel function, referred to as the *social interaction function* or SIF. The SIF formally defines the relationship between two individuals based on spatial proximity. For example it has been shown that many social interaction patterns obey the Zipf law [Zipf, 2016], where individuals are more likely to interact with others close by rather than far away (a pattern that holds even for online interactions [Spiro et al., 2016]). Here, we use this approach to model a network that represents combination of frequent interactions due to ongoing social ties, and contacts resulting from frequent incidental encounters (e.g., interactions with neighbors and community members).

We follow the protocol of [Butts et al., 2012b, Hipp et al., 2013] to simulate social network data that combines the actual distribution of residents in a city with a pre-specified SIF. We employ the model of [Hipp et al., 2013] with decennial Census data to produce large-scale social networks for 19 cities and counties in the United States – providing a representation of major urban areas in the United States (see SI Appendix). Given these simulated networks, we then implement an individual-level SIR-like framework to examine COVID-19 diffusion. At each moment in time, each individual can be in a *susceptible*, *infected but not infectious*, *infectious*, *deceased*, or *recovered* state. The disease diffuses through the contact network, with currently infectious individuals infecting susceptible neighbors as a continuous time Poisson process with a rate estimated from mortality data (see Appendix A); recovered or deceased individuals are not considered infectious for modeling purposes. Upon infection, an individual's transitions between subsequent states (and into mortality or recovery) are governed by waiting time distributions based on epidemiological data

as described in SI Appendix. To begin each simulated trajectory, we randomly infect 25 individuals, with all others being considered susceptible. Simulation proceeds until no infectious individuals remain. The same technique to generate social networks and diffuse COVID across them is used for San Francisco. The diffusion process specified for San Francisco is also specified using infection and mortality data from the city stratified by age and sex, which are two important determinants of COVID-19 outcomes. More details on the diffusion model for San Francisco can be found in Appendix B.

From the simulated trajectory data, we produce several metrics to assess spatial heterogeneity in disease outcomes. First, we present infection curves for illustrative cities, showing the detailed progress of the infection and its difference from what an SIR model would posit. We also present a map showing the simulated infection times of nodes across the city of San Fransisco. While an SIR model would predict an absence of systematic variation in the infection curves or the peak infection day for different areal units in the same city, geographically realistic models show considerable disparities in infection progress from one neighborhood to another. To quantify the degree of heterogeneity more broadly, we examine spatial variation in outcomes for each of our city networks. We show that large variations in peak infection days across tracts are typical (often spanning weeks or even months), and that overall correlations of within-tract infection trajectories with the aggregate urban trajectory are generally modest (a substantial departure from what would be expected from an SIR model). To better characterize the nature of the disparity in infection times by region, we turn to Cox Proportional-Hazard models, which we use to determine the effect of additional network cohesion on one's hazard of COVID-19 infection

In addition to these relatively abstract metrics, we also examine a simple measure of the potential load on the healthcare system in each city. Given the locations of each hospital in each city, we attribute infections to each hospital using a Voronoi tessellation (i.e., under the simple model that individuals are most likely to be taken to the nearest hospital if they become seriously ill). Examination of the potential hospital demand over time shows substantial differences in load, with some hospitals severely impacted while others have few cases. Finally, we consider the *social exposure* of individuals to COVID-19, by computing the fraction of individuals with a personal contact who is respectively morbid or deceased. Our model shows considerable differences in these metrics over time, revealing that the pandemic can appear very different to those "on the ground" – evaluating its progress by its impact on their own personal contacts – than what would be suggested by aggregate statistics.

2.2.1 Spatial Network Data

Networks are generated using population distributions from the most recent US census in 2010. Network construction followed the same methodology as Butts et. al. [Butts et al., 2012b]. Hospital information was obtained from the Homeland Infrastructure Foundation-Level Data (HIFLD) database [Department of Homeland Security, 2016]. HIFLD is a an initiative that collects geospatial information on critical infrastructure across multiple levels of government. We employ the national-level hospital facility database, which contains locations of hospitals for the 50 US states, Washington D.C., US territories of Puerto Rico, Guam, American Samoa, Northern Mariana Islands, Palau, and Virgin Islands; underlying data are collated from various state departments or federal sources (e.g., Oak Ridge National Laboratory). We employ all hospitals within our 19 target cities, excluding facilities closed since 2019. Latitude/longitude coordinates and capacity information were employed to create a spatial database that includes information on the number of beds in each hospital. The capacity information includes the number of beds that each hospital has available, and can be used to assess strain that a surge in hospitalizations could create.

The dates of the first confirmed case and all the death cases for King County, where Seattle is located, were obtained from The New York Times, based on reports from state and local health agencies [New York Times, 2020]. The death rate was calculated based on population size of each county from the 2018 American Community Survey, and employed to calibrate the infection rate (the only free parameter in the models used here); details are provided in the SI Appendix.

We ran 10 replicates of the COVID-19 diffusion process in each of our 19 cities, seeding with 25 randomly selected infections in each replicate and following the course of the diffusion until no infectious individuals remained. Simulations were performed using a combination of custom scripts for the R statistical computing system [R Core Team, 2020] and the statnet library [Butts, 2008a,b, Handcock et al., 2008a]. Analyses were performed using R.

In addition to the 19 cities that we model for the broad scale simulation study, we also model social networks for San Francisco. We generate 35 realizations of the social contact network in San Francisco. For each of these social networks, we seed 35 infections, which produce 1225 pandemic trajectories. More details on the network generation and the parameterization of the diffusion process can be found in Appendix A for the large scale network study, and Appendix B for the San Francisco case study.

2.3 Results

2.3.1 Smooth Aggregate Infection Trajectories Can Mask Local Outbreak Dynamics

When taken over even moderately sized regions, aggregate infection curves can appear relatively smooth. Although this suggests homogeneous mixing (as assumed e.g. by standard SIR models), appearances can be deceiving. Fig. 2.1 shows typical realizations of infection curves for two cities (Seattle, WA and Washington, DC), showing both the aggregate trajectory (red) and trajectories within individual Census tracts (black). While the infection curves in both cases are relatively smooth, and suggestive of a fairly simple process involving a sharp early onset followed by an initially sharp but mildly slowing decline in infections, within-tract trajectories tell a different story. Instead of one common curve, we see that tracts vary wildly in onset time and curve width, with some tracts showing peaks weeks or months after the initial aggregate spike has passed.

The cases of Fig. 2.1 are emblematic of a more systematic phenomenon: the progress of the infection within any given areal unit often has relatively little relationship to its progress in the city as a whole. The bottom panels of Fig. 2.1 assess this phenomenon over our entire sample, using two different consensus metrics. First, we simply compute the correlation between the infection curves in each pair of tracts (assessed at daily resolution), taking the mean for each tract of its correlation with all other tracts within the city; if the progress of the infection were uniform across the city, the mean correlations would be large and positive. Second, we provide a more direct assessment of the extent to which the set of infection curves can be summarized by a common pattern by taking the variance on the first principal component of the correlation matrix generated from the tract level correlations discussed immediately above. As before, where different parts of the city experience similar patterns of growth and decline in infections, we expect the dimension of greatest shared variance to account for the overwhelming majority of variation in infection rates. Contrary to these expectations, however, Fig. 2.1 shows that there is little coherence in tract-level infection patterns. Mean correlations of local infection curves across tracts typically range from approximately 0 to 0.5, with a mean of approximately 0.2, indicating very little correspondence between infection timing in one tract and that of another. The principal component analysis tells a similar story: overall, we see that first component accounts for relatively little of the total variance in trajectories, with on average only around 35% of variation in infection curves lying on the first principal component (and no observed case of the first component accounting for more than 60% of the variance). Interestingly, this variation is not explained by

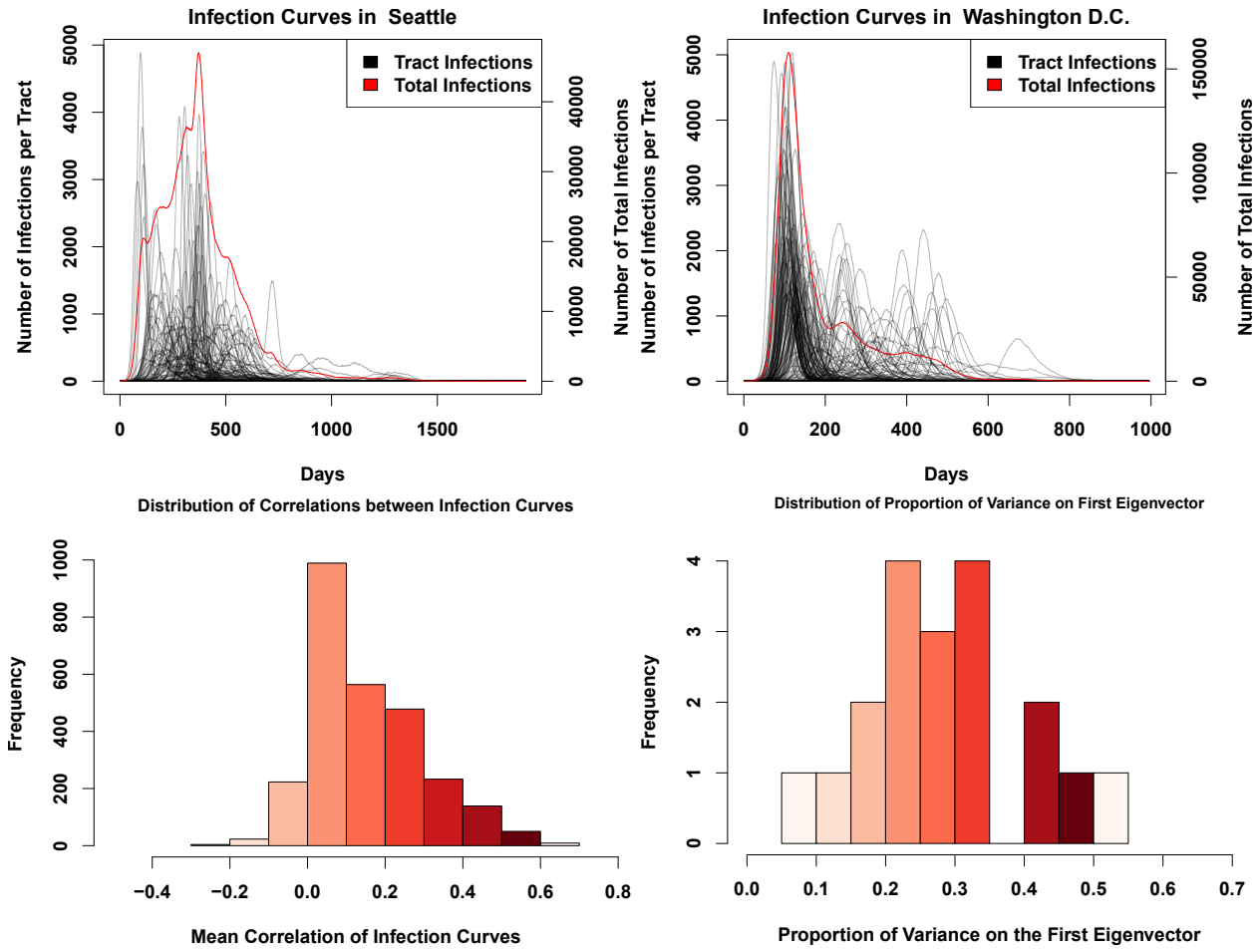


Figure 2.1: (Top Left) Infection curves for Seattle, WA. The red line is the curve for the whole city, while the black lines are the the infection curves for each tract in the city. While the Red curve is relatively smooth, this smoothness hides a significant amount of heterogeneity in the timing of the infection curves for each census tract. (Top Right) Infection Curves for Washington D.C. As with Seattle, the city-level curve conceals considerable spatial variability in the infection's progress. (Bottom Left) Histogram showing the mean pairwise correlation of infection curves for each tract within each city, across our entire sample. The infection curve in any given tract is likely to have a correlation of only around 0.2 with any other tract in the city. This histogram includes a single datapoint for each tract in the sample. (Bottom Right) Histogram of variance accounted for by the principal component of the standardized tract-level curve set. None of the principal components account for more than 60% of the variance, with most accounting for around only 35% of the total variance. The datapoints included here include a single amount of variance explained for each city.

time required for the diffusion process to reach each tract (see SI Appendix figure S4), in contrast to the hypothesized importance of similar delays in a cross-national context [Brockmann and Helbing, 2013]. This confirms that *local infection curves are consistently distinct*, with behavior that is only weakly related to infections in the city as a whole. This is a substantially different scenario than what is commonly assumed in traditional SIR models.

Spatially, we also observe a significant amount of heterogeneity in infection times. We examine individual

level infection times in the model to better understand the effects of spatial heterogeneity. Figure 2.9B describes the pattern of infection timings across San Francisco. Infections time vary significantly across the city, with the majority of early infections taking place in the downtown regions of the city. Areas like Richmond and the Mission are spared early infections, but have high infection rates across the middle of the pandemic trajectory. Meanwhile, the Twin Peaks and Diamond Heights regions are spared many infections across the entire simulated length of the pandemic. These spatial patterns show that there is significant clustering in the patterns of COVID-19 infections, for which we propose and examine potential mechanisms for in Section 2.3.4.

2.3.2 Heterogeneous Impact Timing May Affect Hospital Load

Variation in the timing of COVID-19 impacts across the urban landscape has potential ramifications for healthcare delivery, creating unequally distributed loads that overburden some providers while leaving others with excess resources. To obtain a sense of how spatial heterogeneity in the infection curve could potentially impact hospitals, we employ a simple "catchment" model in which seriously ill patients are taken to the nearest hospital, subsequently recovering and/or dying as assumed throughout our modeling framework. Based on prior estimates [Stokes et al., 2020], we assume that 14% of all infections are severe enough to require hospitalization (robustness to alternative rate estimates shown in the SI Appendix). While hospitals draw from (and hence average across) areas that are larger than tracts, the heterogeneity shown in Fig. 2.1 suggests the potential for substantial differences in hospital load over time. Indeed, our models suggest that such differences will occur. Fig. 2.2 shows the number of patients arriving at each hospital in Seattle and Washington, DC (respectively) during a typical simulation trajectory. While some hospitals do have demand curves that mirror the city's overall infection curve, others show very different patterns of demand. In particular, some hospitals experience relatively little demand in the early months of the pandemic, only to be hit hard when infections in the city as a whole are winding down.

Just as hospital load varies, hospital capacities vary as well. As a simple measure of strain on hospital resources, we consider the difference between the number of COVID-19 hospitalizations and the total capacity of the hospital (in beds), truncating at zero when demand outstrips supply. (For ease of interpretation as a measure of strain, we take the difference such that higher values indicate fewer available beds.) Using data on hospital locations and capacities, we show in Fig. 2.2 strain on all hospitals in Seattle and Washington, D.C. (respectively) during a typical infection trajectory. While some hospitals are hardest hit early on (as would be expected from the aggregate infection curve), others do not peak for several months. Likewise,

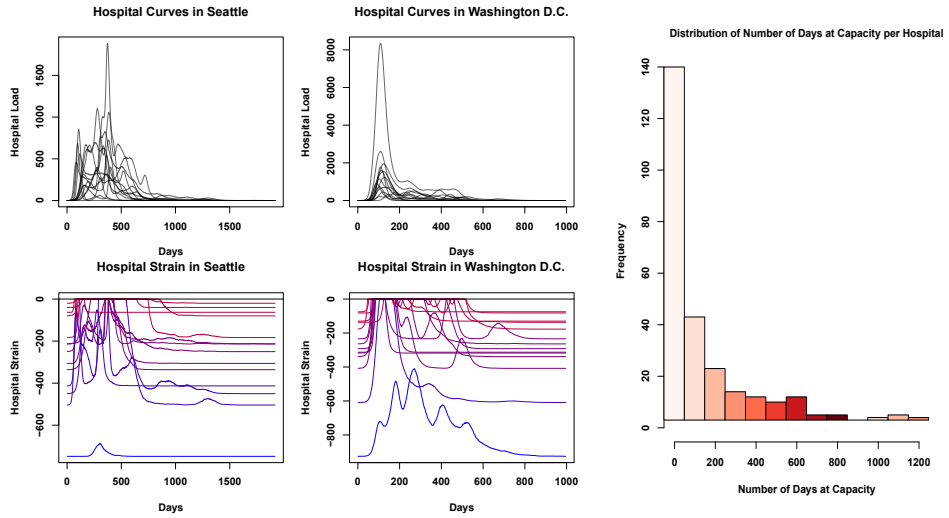


Figure 2.2: (Top Left) Numbers of infections attributed to each hospital in the city of Seattle, with each curve representing a different hospital. Hospital peak demand times vary markedly, with some getting the majority of their hospitalizations before day 100, and others peaking almost a year into the pandemic. (Top Middle) Hospitalizations in Washington, DC. As in Seattle, the each hospital has a unique demand trajectory, with some hospitals not getting their peak of infections until more than a year after the infection begins. (Bottom Right) Hospital strain in Seattle, WA. Values closer to zero indicate that hospitals are more strained and have fewer open beds, while lower values suggest more resources are available; color varies from blue (low average strain) to red (high average strain). Much like the number of infections, there is a high degree of heterogeneity present here, with hospitals freeing up resources at different points across the first year of the pandemic. (Bottom Middle) Hospital strain for Washington D.C. Most hospitals get overwhelmed in the first 25 days of the pandemic, but then are able to recover at different times, usually within the second hundred days of the pandemic; some, however, are hit hard by a second wave, and others remain overwhelmed for several months (Right) Marginal distribution of number of days without available beds, for all hospitals in our sample. While most hospitals will have only brief periods of overload, some will be at or over capacity for the entire pandemic, potentially several years.

hospitals proximate to areas of the city with very different infection trajectories experience natural "curve flattening," with a more distributed load, while those that happen to draw from positively correlated areas experience very sharp increases and declines in demand. These conditions in some cases combine to keep hospitals well under capacity for the duration of the pandemic, while others are overloaded for long stretches of time. These marked differences in strain for hospitals within the same city highlight the potentially complex consequences of heterogeneous diffusion for healthcare providers.

Looking across cities, we see the same high-variability patterns as observed in Seattle and Washington. In particular, we note that local variation in disease timing leads to a heavy-tailed distribution for the duration at which hospitals will be at capacity. Fig. 2.2 shows the marginal distribution of hospital overload periods (defined as total number of days at capacity during the pandemic), over the entire sample. While the most common outcome is for hospitals to be stressed for a brief period (not always to the breaking point), a significant fraction of hospitals end up being overloaded for months - or even, in a small fraction of cases, nearly the whole duration of the pandemic.

It should be reiterated that the hospital load model used here is extremely simplified, and that we are employing a no-mitigation scenario. However, these results quite graphically demonstrate that the importance of curve-flattening interventions does not abate once geographical factors are taken into account. On the other hand, these results suggest that differences in hospital load may be substantially more profound than would be anticipated from uniform mixing models, creating logistical challenges and possibly exacerbating existing differences in resource levels across hospitals. At the same time, such heterogeneity implies that resource sharing and patient transfer arrangements could prove more effective as load-management strategies than would be suggested by spatially homogeneous models, as hospitals are predicted to vary considerably in the timing of patient demand.

2.3.3 Social Exposures to Morbidity and Mortality Vary by Location

In addition to healthcare strain, the *subjective experience* of the pandemic will potentially differ for individuals residing in different locations. In particular, social exposures to outcomes such as morbidity or mortality may shape individuals' understandings of the risks posed by COVID-19, and their willingness to undertake protective actions to combat infection. Such exposures may furthermore act as stressors, with potential implications for physical and/or mental health. As a simple measure of social exposure, we consider the question of whether a focal individual (ego) either has experienced a negative outcome themselves, or has at

least one personal contact (alter) who has experienced the outcome in question. (Given the highly salient nature of COVID-19 morbidity and mortality, we focus on the transition to first exposure rather than e.g. the total number of such exposures, as the first exposure is likely to have the greatest impact on ego's assessment of the potential severity of the disease.)

To examine how social exposure varies by location, we compute the fraction of individuals in each tract who are socially exposed to (respectively) morbidity or mortality. Fig. 2.7 shows these proportions for Baltimore, MD, over the course of the pandemic. As with other outcomes examined here, we see considerable variation in timing, with many tracts seeing a rapid increase in exposure to infections, while others go for weeks or months with relatively few persons having a personal contact with the disease. Another notable axis of variation is sharpness. In many tracts, the fraction of individuals with at least one morbid contact transitions from near zero to near one within a matter of days, creating an extremely sharp social transition between the "pre-exposure world" (in which almost no one present knows someone with the illness) to a "post-exposure world" in which almost everyone knows someone with the illness). By contrast, other tracts show a much more gradual increase (sometimes punctuated by jumps), as more and more individuals come to know someone with the disease. In a few tracts that are never hit hard by the pandemic, few people ever have an infected alter; residents of these areas obviously have a very different experience than those of high-prevalence tracts. These distinctions are even more stark for mortality, which takes longer to manifest and which does so much more unevenly. Tracts vary greatly in the fraction of individuals who ultimately lose a personal contact to the disease, and in the rapidity with which that fraction is reached. In many cases, it may take a year or more for this quantity to be realized; until that point, many residents may be skeptical to the notion that the pandemic poses a great risk to them personally.

By way of assessing the milieu within each tract, it is useful to consider the "cross-over" point at which at least half of the residents who will be socially exposed in a given tract have been socially exposed to either COVID-19 morbidity or mortality. Fig. 2.5 and Fig 2.6 map these values for Baltimore, MD. It is immediately apparent that social exposures are more strongly spatially autocorrelated than other outcomes considered here, due to the presence of long-range ties within individuals' personal networks. Even so, however, we see strong spatial differentiation, with residents in the urban core being exposed to both morbidity and mortality much more quickly than those on the periphery. This suggests that the social experience of the pandemic will be quite different for those in city centers compared to those in more outlying areas, with the latter taking far longer to be exposed to serious consequences of COVID-19. This may manifest in differences in willingness to adopt protective actions, with those in the urban core being more highly motivated to take

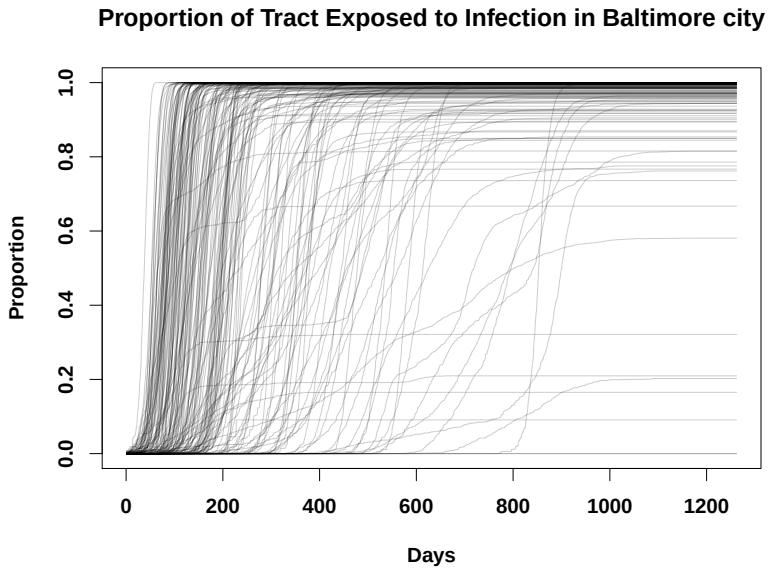


Figure 2.3: (Left) Trajectories showing the fraction of people in each tract in Baltimore who have an infected person in their personal network across time. We see a large degree of spatial heterogeneity, as some tracts are more insulated from others in terms of social exposure. However, by the end of the pandemic, most people across all tracts have been exposed to someone who has had the disease.

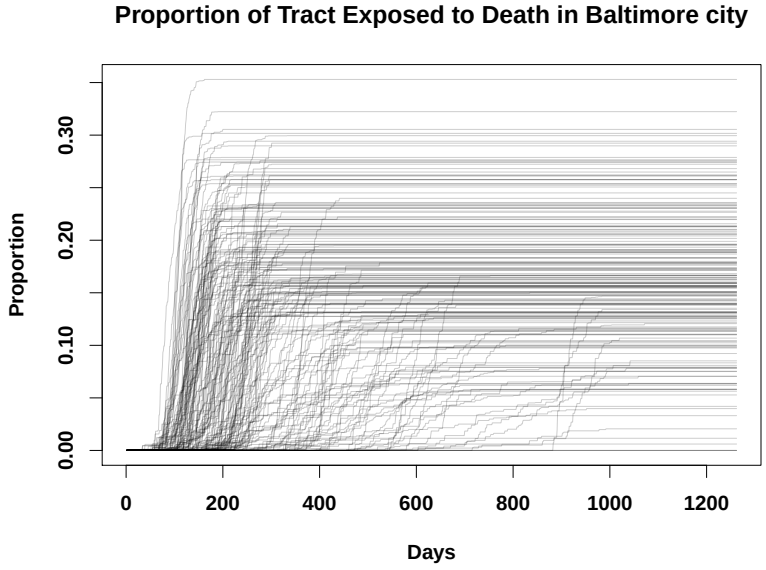


Figure 2.4: (Right) The fraction of persons in each tract who have an alter who died from COVID-19 in their personal network. On average, only around 20-30% of people in any given tract know someone who died by the end of the pandemic, though this varies widely across tracts.

action (and perhaps resistant to rhetoric downplaying the severity of the disease) than those on the outskirts of the city.

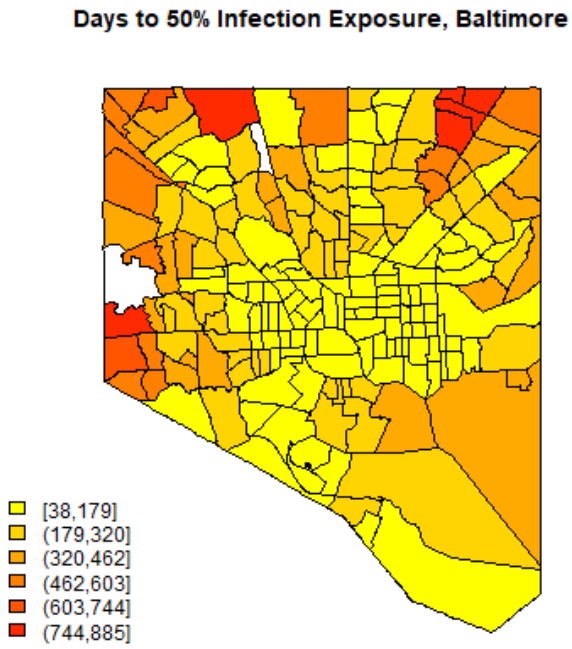


Figure 2.5: Chloropleth showing the time for half of those in each tract to be socially exposed to COVID-19 morbidity in Baltimore, MD. The central parts of the city are exposed far sooner than the northwestern part of the city.

2.3.4 The Mechanistic Drivers of COVID-19 Spread in Social Networks

Fig. 2.7 shows, wild-type SARS-CoV-2 does not diffuse readily through linear "infection chains" with multiple intermediates; even when multiple, parallel chains connect two individuals, many chains are required to achieve a large infection risk. By contrast, SARS-CoV-2 spreads extremely well through *cohesive subgroups*, where multiple, redundant ties provide numerous avenues for infection to occur. Being connected to an infective by shared membership in even a fairly small cohesive group results in a dramatic increase in infection risk, due to the factorial increase in the number of potential infection paths with group size. For example, an otherwise isolated susceptible linked to an infective via a clique of only 6 individuals has a 50% probability of becoming infected; to reach the same infection probability by connection with independent paths of the type shown in Fig. 2.7 would require maintaining 38 contacts involving 76 intermediaries. This suggests that *small differences in social cohesion can lead to large disparities in infection risk* for wild-type SARS-CoV-2, much as small differences in partnership concurrency have been shown to drive disparities in HIV risk [Morris et al., 2010].

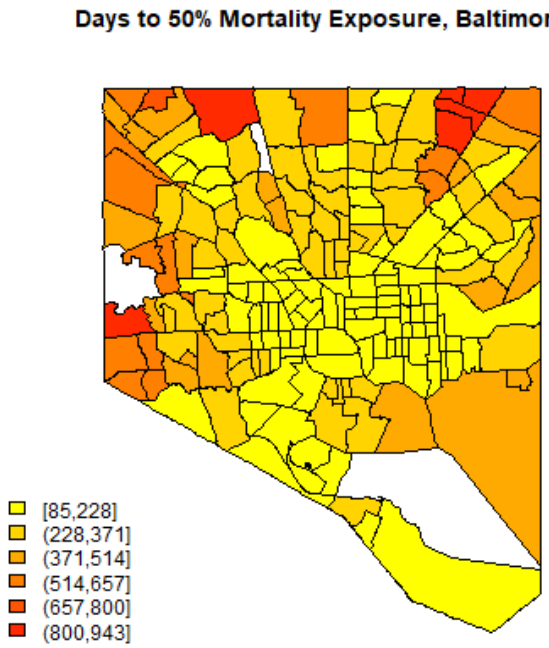


Figure 2.6: Chloropleth showing the time for half of those in each tract to be socially exposed to COVID-19 mortality. Central Baltimore is exposed to deaths in personal networks far sooner than the more outlying areas of the city.

To determine whether these network effects would be expected to manifest under realistic conditions, we employ the above model [Thomas et al., 2020] to study early pandemic infection hazards in the city of San Francisco, CA, a major city with a diverse population that suffered significant disparities in pandemic outcomes. This case study examines the period before March 24, 2020, one week after infection data became available for the four major racial/ethnic groups; by this time, the infection was already spreading throughout the city, and significant racial and ethnic disparities in incidence had emerged. The observed patterns of disparity are typical of what would be expected given the underlying network process with disparities in infection risks being greatly enhanced by differences in social cohesion. As we further show through simulation, these differences are expected to be geographically correlated, leading to a high risk "flood plain" that is particularly exposed to infection, and metaphorical "high ground" that is relatively protected.

Infection Outcomes: We simulate 1225 infection trajectories ("pandemic histories") for the city of San Francisco (see Section 2.2 and Appendix B) covering the period up to March 24, 2020. Fig. 2.8B shows

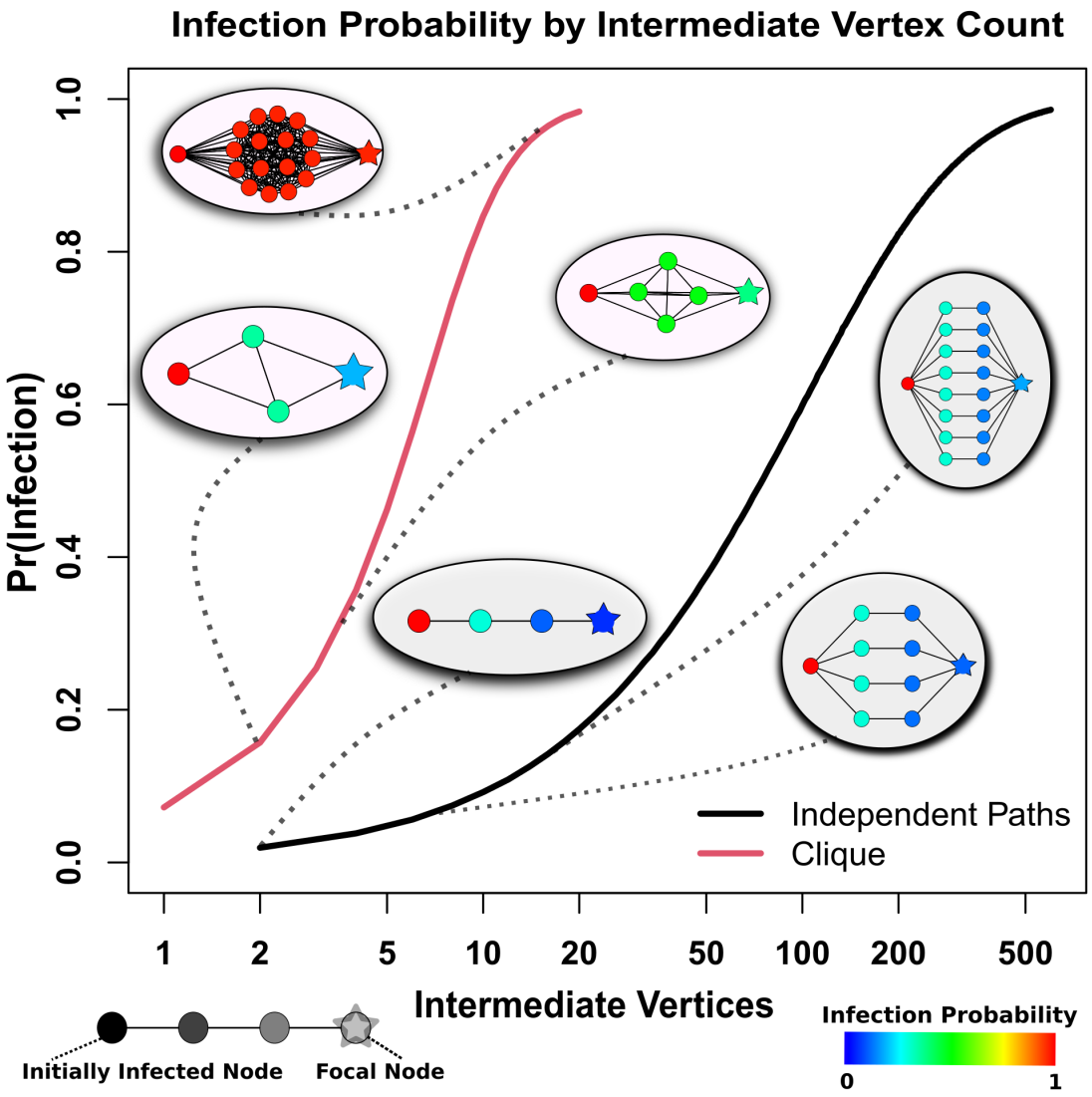


Figure 2.7: Probability of diffusion from an infected (left) to uninfected (right) individual bridged by intermediaries arranged in cliques (red curve) versus independent paths (black curve). Co-membership in a cohesive subgroup fields infection risks that climb sharply with the number of intermediaries, while much larger numbers of intermediaries are required to obtain the same risk in the case of independent paths.

the resulting distribution of early infection disparities by demographic group (Hispanic (H), Non-Hispanic Black (B), Non-Hispanic White (W), and Non-Hispanic Asian (A)) on March 24, 2020 of the simulation. Because outbreaks can vary greatly in size and timing, early period disparities can and do vary by trajectory. However, we see that Hispanics are hardest-hit in the majority of cases, typically followed by Blacks and then Asians. Non-Hispanic Whites are very rarely the hardest-hit, and are often (but not always) the group with the lowest early incidence; we note more variability in the identity of the least-hit group, as this outcome

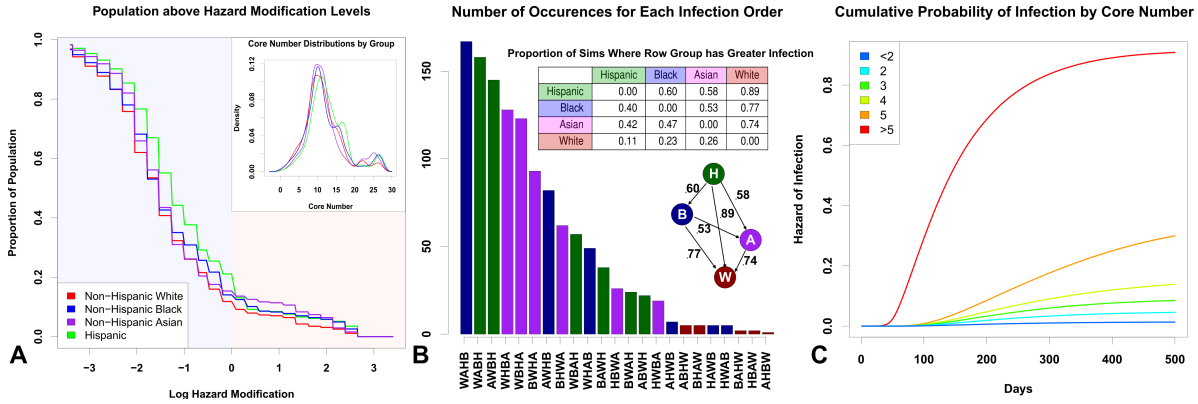


Figure 2.8: (A)(Inset) Distribution of core numbers for each ethnoracial group in the SF model; small differences in core numbers are sufficient to drive large differences in risk. (Main) Proportion of each population that lives "below" a given point on the floodplain (higher risk), denoted by its log hazard modification. The Non-Hispanic White population is consistently present on the higher parts of the floodplain, with the Non-Hispanic Asian population also being present in the middle of the floodplain. The lower parts of the floodplain are heavily occupied by Non-Hispanic Black and Hispanic populations. (B) Distribution of qualitative outcomes in simulation on March 24, where x-axis labels correspond to group labels in order of infection rates, from lowest (bottom) to highest (top) prevalence. Columns are colored corresponding to the group with highest prevalence. The third column (order AWBH) corresponds to the observed pattern from San Francisco. (Inset) The proportion of times each row group has a greater infection rate than the column group across all simulations. The Hispanic population consistently has the highest infection rates, followed on average by the Black population, the Asian population, and the non-Hispanic White population. (C) Cumulative probability of infection by core number from simulated networks. Higher core numbers indicate greater levels of local cohesion, which substantially increases one's hazard of infection. The bicomponent, where core number is equal to 2 does not seem to drive infection patterns, as some prior literature suggests Moody, James and Adams, Jimi and Morris, Martina [2017].

is sensitive to chance events (i.e., where early outbreaks occur). The observed pattern based on official data San Francisco Department of Public Health [2021b] is the third-most common pattern that would be expected, and hence fairly typical of what would be expected given the contact process.

Cohesion Drives Infection Hazard: Fig. 2.7 shows the risk-enhancing effect of cohesion in isolated subnetworks; this effect generalizes to more realistic scenarios. A Cox proportional hazards model of infection hazard by core number (a common measure of embeddedness in cohesive groups) confirms a large risk enhancement for local cohesion, with persons in cohesive subgroups facing dramatically higher infection risk over time (Fig. 2.8C); in particular, each unit increment in core number increases infection hazard by apx 30%. Different demographic groups have slightly different levels of cohesion (Fig. 2.8A, inset). The difference in mean core number between the most cohesive group (Hispanic) and the least (Non-Hispanic White) is 1.5, translating to an approximately 50% mean risk enhancement; while risk levels vary within all groups, a 9.3% higher share of Hispanic versus Non-Hispanic White population has greater than average risk (Fig. 2.8A). Differences in local social cohesion thus provide an important structural basis for disparities in

early pandemic outcomes between groups.

Spatial Correlation of Cohesion Produces a Network "Floodplain:" Contact network cohesion is spatially correlated, producing areas with higher than average membership in cohesive subgroups, and hence elevated mean risk. Fig. 2.9A shows the mean infection hazard modifier (net of global average) for each U.S. Census block in San Francisco, based on the distribution of cohesion scores (core numbers). Cyan and green areas are epidemiological "high ground" where lower levels of local cohesion reduce mean risk, while red and orange areas are epidemiological "floodplains" where high cohesion leads to enhanced local risk. These cohesion-driven patterns are well-correlated with the overall rate of infections, as illustrated by the mean inverse infection time across the city (Fig. 2.9B). Spatial segregation in housing places some groups in harm's way, increasing disparities in incidence during the initial outbreak.

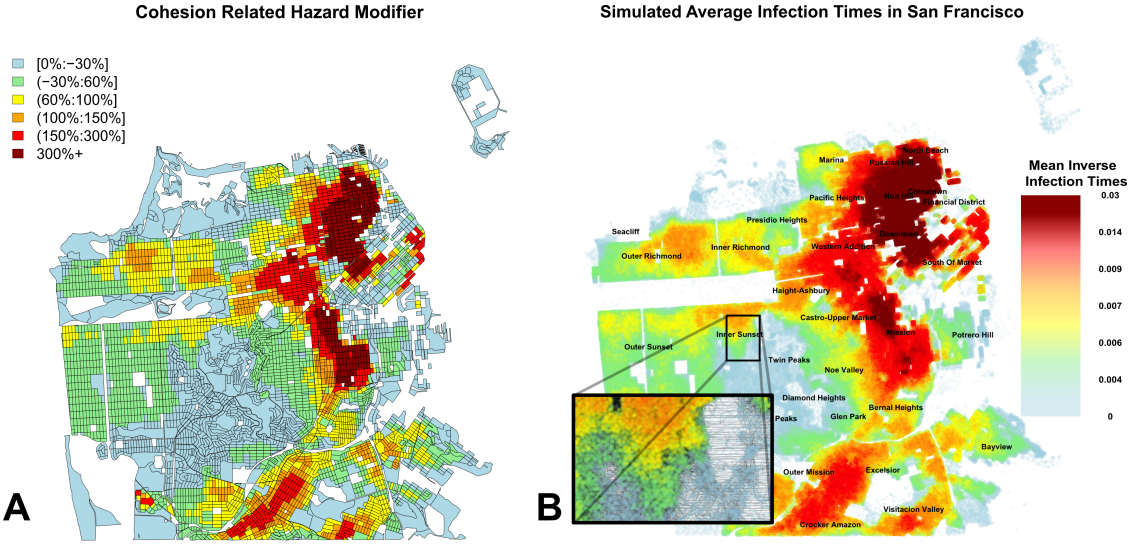


Figure 2.9: (A) Average deviation from the mean hazard attributable to core number, across San Francisco. Risk enhancement is spatially correlated, with significant risk downtown and much lower risk near the central part of the city. These hazards form a "floodplain," where some areas are more dangerous than others. (B) Simulated infection times across San Francisco, averaged across 35 simulations. The patterns of infections match the expected hazard modifications in the left panel. The inset shows the structure of the social network in the Inner Sunset neighborhood.

2.4 Discussion

Our simulation results all underscore the potential effects of local spatial heterogeneity on disease spread. The spatial heterogeneity driving these results occurs on a very small scale (i.e., Census blocks), operating well below the level of the city a whole. As the infection spreads, relatively small differences in local

network connectivity and the prevalence of bridging ties driven by uneven population distribution can lead to substantial differences in infection timing and severity, leading different areas in each city to have vastly different experiences of the pandemic. Resources will be utilized differently in different areas, as some areas will experience the bulk of their infections far later than others, and the subjective experience of a given individual regarding the pandemic threat may differ substantially from someone in another area. These behaviors are in striking contrast to what is assumed by models based on the assumption of spatially homogeneous mixing, which posit uniform progress of the infection within local areas.

As noted at the outset, our model is based on a no-mitigation scenario, and is not intended to capture the impact of social distancing. While distancing measures by definition limit transmission rates - and will hence slow diffusion - contacts occurring through spatially correlated networks like those modeled here are still likely to show patterns of heterogeneity like those described. One notable observation from our simulations is the long outbreak delay that some census tracts experience, even in the absence of social distancing. This would suggest that relaxation of mitigation measures leading to a resumption of "normal" diffusion may initially appear to have few negative effects, only to lead to deadly outbreaks weeks or months later. Public health messaging may need to stress that apparent lulls in disease progress are not necessarily indicators that the threat has subsided, and that areas "passed over" by past outbreaks could be impacted at any time.

Finally, we stress that conventional diffusion models using locally homogeneous mixing have been of considerable value in both pandemic planning and scenario evaluation. Our findings should not be taken as an argument against the use of such models. However, the observation that incorporating geographical heterogeneity in contact rates leads to radically different local behavior would seem to suggest that there is value in including such effects in models intended to capture outcomes at the city or county level. Since these are the scales on which decisions regarding infrastructure management, healthcare logistics, and other policies are often made, improved geographical realism could potentially have a substantial impact on our ability to reduce lives lost to the COVID-19 pandemic.

We also find that the mere presence of connecting paths is not sufficient for rapid diffusion of a disease like wild type SARS-CoV-2: infection of contacts is rare enough to require considerable redundancy for transmission to occur. Cohesion greatly increases the number of potential infection pathways, rendering an otherwise relatively "opaque" network "transparent" to disease transmission. The uneven distribution of cohesive subgroups in large networks and their much greater permeability helps to explain the "bursty" nature of SARS-CoV-2 diffusion, with slow diffusion through less cohesive parts of the network punctuated by rapid outbreaks in cohesive groups [Thomas et al., 2020, Wong and Collins, 2020]. Ironically, social

cohesion has long been viewed as a community asset, particularly with respect to community resilience following disasters or other sources of social disruption [Fan et al., 2020, Townshend et al., 2015, Cinner, JE and Lau, JD and Bauman, AG and Feary, DA and Januchowski-Hartley, FA and Rojas, CA and Barnes, ML and Bergseth, BJ and Shum, E and Lahari, R and others, 2019]; in the context of an infection like SARS-CoV-2, this same cohesion can act as an epidemiological risk factor. Local cohesion varies by location, with some parts of the San Francisco network having higher local cohesion than others. Combined with high levels of residential segregation, these differences can in turn produce disparities in infection hazard by race and ethnicity. In San Francisco, we find that Black and Hispanic populations are expected to have the highest infection rates in the early pandemic, followed by the Asian population and the White Non-Hispanic population. Our models suggest that the exact evolution of infection rates is somewhat contingent on chance events, and multiple scenarios are possible based on which subgroups are hit first; however, some scenarios are much more likely than others, with the observed pattern of infection in the early pandemic being one of those predicted to be most likely to occur. Greater attention to cohesion as a risk factor - particularly given its spatial correlation - may help to prioritize warning messages or interventions for high-risk groups when outbreaks of a potentially serious disease are first detected.

Chapter 3

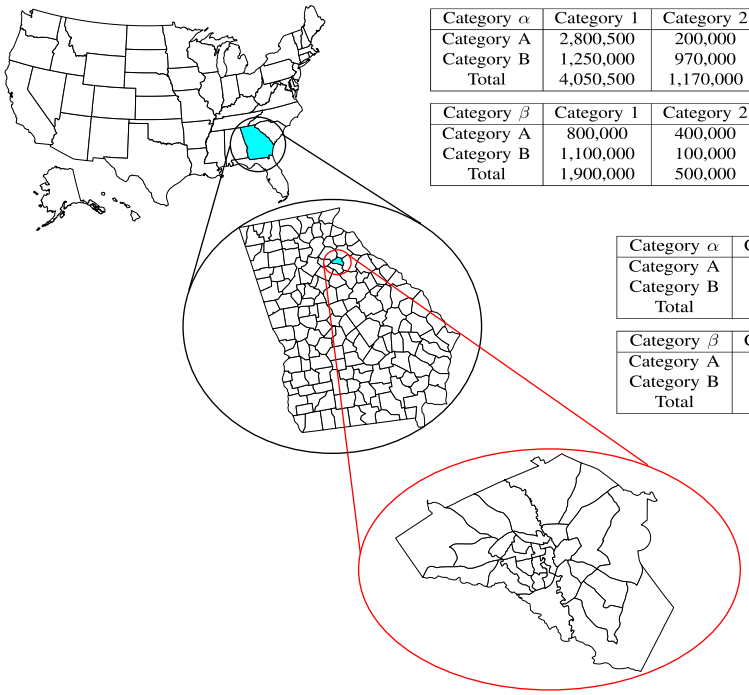
Marginal-preserving Imputation of Three-way Array Data in Nested Structures, with Application to Small Areal Units

3.1 Introduction

Many data sources, including the U.S. Census and organizations using Google's S2 projection system [1], provide geospatial population data organized into a nested hierarchy of areal units. In such hierarchical structures, each areal unit at a given level can be expressed as the union of a set of units at the level below, in turn being part of a single parent; each level is hence a spatial partition of the region of interest (Fig. 3.1). Many sociological questions involve the cross-tabulation of population properties within such units with other quantities (e.g., environmental, ecological, political, economic, or other variables that vary across regions). With the advent of increasingly well-developed spatial data sets [Rose et al., 2021, Facebook Connectivity Lab et al., 2016], performing such analyses at increasingly fine geographical resolution is of

[1] The S2 geometries use nested areal units on the sphere, and can be used to describe spatial relationships across Earth based on sets of locations and attributes.

substantial interest [Thomas et al., 2020].



Category α	Category 1	Category 2	Total
Category A	2,800,500	200,000	3,000,500
Category B	1,250,000	970,000	2,220,000
Total	4,050,500	1,170,000	–

Category β	Category 1	Category 2	Total
Category A	800,000	400,000	1,200,000
Category B	1,100,000	100,000	2,220,000
Total	1,900,000	500,000	–

Category α	Category 1	Category 2	Total
Category A	100,000	15,000	115,000
Category B	120,000	80,000	200,000
Total	220,500	95,000	–

Category β	Category 1	Category 2	Total
Category A	75,000	45,000	120,000
Category B	120,000	15,000	135,000
Total	195,000	60,000	–

Category α	Category 1	Category 2	Total
Category A	?	?	213
Category B	?	?	104
Total	180	137	–

Category β	Category 1	Category 2	Total
Category A	?	?	115
Category B	?	?	200
Total	195	60	–

Figure 3.1: An example of a hierarchical data structure on areal units, using the US Census areal unit hierarchy. We are interested in tabulating population with respect to three hypothetical dimensions, represented by respective category sets $\{\alpha, \beta\}$, $\{1, 2\}$, and $\{A, B\}$. At higher levels in the hierarchy, we may have complete areal unit data with respect to all categories; but for small units, we may have only marginal information (third table). By combining marginal information at fine-grained units with associations observed in the more complete parent data, we impute cross-tabulations for the fine-grained units.

In practice, however, such fine-grained analyses can still encounter problems of data availability. For instance, while detailed Census data is publicly released at higher levels of the Census geography (e.g., counties), incomplete data is released at smaller geographical scales (e.g., blocks and block groups). This issue is not unique to the U.S. Census: releasing fully detailed information at fine scale poses challenges of acquisition cost (there are vastly more small areal units than large ones), availability (key variables may not be obtained at all scales), distribution and maintenance costs, and privacy considerations. Where information for smaller units is available, it is often available only marginally (i.e. summed across all values of a covariate), without the cross-tabulation needed to study many demographic processes. For instance, we may know how many individuals reside in a given unit by race, by ethnicity, and by gender, but we may not know how many White Hispanic women reside there. Raising our level of analysis to the smallest unit with complete tabulation may resolve this difficulty, but at cost of "blurring" spatial heterogeneity. Particularly when studying phenomena that occur on small scales - e.g., neighborhood interactions, exposure to crime or other events, or immediate access to local amenities - this causes problems for analysis.

While there is no perfect substitute for complete data, the presence of incompletely tabulated data suggests the viability of imputation strategies: even one-way marginals can be powerfully constraining, and two-way marginals even more so. When marginals can be combined with information on correlations from higher-order units with complete data, it may be possible to accurately estimate the local tabulation in a way that preserves all known quantities. This preserves spatial heterogeneity and permits fine-grained analysis, while also making use of more complete information where available. Surprisingly, this approach to the multi-way areal unit imputation problem appears to have been overlooked in prior literature, though we draw on a number of related developments in our work (as described below).

In this paper, we introduce a method for imputing cross-tabulated count data organized into a nested system of hierarchical bins, that is highly parallelizable and hence applicable to large systems (prominently including the U.S. Census). We focus on the case of data that is cross-tabulated with up to three different discrete features, each of which may take on a number of values (i.e., a three-way crosstab); our approach combines lower-order information on marginals from the focal bin with more complete, higher-order marginals from the bin's parent to impute the full multi-way array. We can verifiably preserve all available information on the focal bin (assuming that such data is consistent), while approximating higher-order information to the extent possible given low-order constraints. Our technique also allows either point estimation, or simulation of draws from the conditional maximum entropy distribution of the target array given the observed data constraints, supporting use cases such as multiple imputation that is capable of offering consistent uncertainty measures [Rubin, 1996]. As an illustration of the method, we apply our approach to imputation of small areal unit data using the 2010 U.S. Decennial census, demonstrating how it enables fine-grained ecological analysis (here, of differences in exposure to crime) despite data constraints.

3.2 Prior Work

As noted, the specific problem of marginal-preserving multi-way count-data imputation from combined marginal and hierarchical information seems not to have been addressed in prior work. However, a number of related problems have been studied, solutions to which inform our own approach. By way of background, we thus begin by reviewing related results on small areal unit estimation and imputation for three-way cross-tabs, both of which set the stage for our work.

3.2.1 Small Areal Unit Estimation

In its more general context, the problem of inferring characteristics of (usually small) geographical regions is known as the small area estimation (SAE) problem. This is a challenge that arises in many different fields, and work on SAE has likewise bridged a number of disciplines, including but not limited to sociology, demography [Morrison, 1971], and statistics [Graham et al., 2009, Bunea and Besag, 2000]. While small areal unit estimation often deals with estimation of population demographics *per se*, some work goes beyond this to examine covariates such as poverty or disease [Pfeffermann et al., 2013, Molina and Rao, 2010]. As noted in this paper, these techniques are also applicable to examinations of crime exposure in a population. There are numerous strategies for this problem, ranging from simpler strategies such as uniform imputation (completely uninformed at the small geographic unit) or spatial smoothing techniques such as kriging that attempt to flexibly exploit spatial autocorrelation across units [Bennett et al., 1984, Mooney et al., 2020], to more informed model-based approaches [Cohen and Di Zhang, 1988, Steinberg, 1979]. Related to this work, some literature has specifically examined work on maintaining structural constraints and the use of model assisted approaches [Espuny-Pujol et al., 2018, Luna et al., 2015, Moretti and Whitworth, 2020].

In the field of criminology, the interest in estimating models in which crime is an outcome measure in increasingly small geographic units has resulted in a need for small area estimation. Whereas some scholars have simply utilized a uniform imputation strategy to assign data from a larger geographic unit to a smaller unit, another strategy occasionally utilized synthetic estimation for ecological inference [Boessen and Hipp, 2015]. This strategy requires the assumption that the relationships between variables in the larger geographic units are the same as the relationships within the smaller geographic subunits.

Surprisingly few studies in criminology or sociology have explored the exposure to crime of different demographic groups. Arguably, this state of affairs is due to the difficulty of obtaining crime data at a more granular scale. For example, one study measured the context of small suburban communities (defined as population less than 10,000) in assessing the exposure to crime [Alba et al., 1994]. Another study in Cleveland aggregated census tracts to "neighborhoods," and thus an even larger geographic unit [Logan and Stults, 1999]. Yet another study measured the context as police precincts in New York City, which are larger yet given that there were 75 at the time of the study in a city with over 7 million residents [McNulty, 1999]. The challenge is that such units may be too large to capture the environment of a specific person, or group of people. This issue arises in the context of other social exposures, as well; for instance, Thomas et al. [2020, 2022b] provide evidence that both infection hazards and social exposure to others' morbidity and

mortality in the early COVID-19 pandemic was affected by local variation in network structure influenced by housing and demographic factors at or below the block scale. Such differences in exposure may affect not only immediate health outcomes, but also responsiveness to public health interventions, with consequences for both policy effectiveness and health disparities.

The SAE problem is computationally challenging, both because it often involves discrete optimization (e.g., for population counts) and because SAE solutions are often intended to be used at scale: given that e.g., the U.S. has over 8 million Census blocks, and there are over 1.6 billion level 14 S2 cells worldwide (each about 500m across), efficiency can be a significant concern. As such, work on this problem has spurred a range of computational advances, from algorithms to actually perform estimation *per se* [Graham et al., 2009, Vermunt et al., 2008] to the evaluation of produced results [Pfeffermann and Correa, 2012]. Many of the more statistically principled algorithms derive from the literature on hierarchical Bayesian modeling, which provides numerous conceptual and statistical tools for flexible estimation and incomplete pooling of information across units [King et al., 1999]. While these frameworks often require significant computational resources, as the Markov Chain Monte Carlo (MCMC) algorithms required for fitting and simulating draws from such models [Rosen et al., 2001] are very computationally intensive, the algorithms enable estimations of more complex problems where the joint probability functions are not in closed form.

In this context, our work contributes to the SAE literature by implementing an algorithm for Small Areal Unit Estimation that can produce imputed cross-classification data for areal units that satisfy a complex constraint structure (guaranteeing that imputations exactly preserve one and two-way marginal totals, and are integer-valued), while also including information from higher-order units. Our technique draws on the statistical strengths of the SAE literature by leveraging a hierarchical model, extending the work of [Bunea and Besag, 2000] by including additional information about the composition of larger areal units in the imputation process.

3.2.2 Imputation for Cross-tabulated Count Data

Apart from the SAE problem, our work is also related to the general problem of imputation for cross-tabulated count data. In general form, this problem involves a target matrix $N \in \mathbb{N}^{c_1 \times \cdots \times c_d}$ (with c_i being the size of the ith dimension, $\mathbb{N}$ the natural numbers, and $\times$ the Cartesian product), from which only a subset of cells (or, in many cases, marginal totals) is observed. The problem is then to produce a matrix $\hat{N}$ that approximates N, while preserving all observed quantities. For the purposes of this paper, we specifically

focus on the three-way case (i.e., $d = 3$ in the above), as this case allows for significant variability for table interiors. Solving the three-way case also provides two way tables via marginal counts. Naively, the most basic option for three-way imputation is to evenly allocate population to each cell in the three-way crosstab. This preserves the 0-way marginal (i.e., the population total), but not other marginals. One-way or two-way marginals can be preserved by a continuous relaxation of the problem in which each cell is given the same value used as the expected value in the corresponding k-way Chi-Square test [McHugh, 2013], but this does not provide an integer solution - obtaining integer solutions that exactly satisfy the marginal constraints can be substantially more difficult [Bunea and Besag, 2000].

Beyond preserving marginal (or other) information in N, one may seek to preserve (or approximate) more general patterns of associations (e.g., correlations among category memberships). Again, the continuous relaxation of this problem is substantially simpler than the exact version, and indeed it has been extensively studied in the context of log-linear models [Clogg and Eliason, 1987]. Log-linear models represent the expected count for each cell in an array as a multiplicative combination of interactions, such that the log expectation has a linear form; expected marginals are easily preserved in this framework by incorporating parameters derived from observed marginals, but higher-order associations between category memberships can also be employed. Although the simplest approaches to inference for log-linear models are based on maximum likelihood estimation under the assumption that counts are conditionally Poisson distributed (exploiting the resulting exponential family structure), Bayesian and other forms of regularized inference [Graham et al., 2009, Vermunt et al., 2008] have also been employed. Log-linear models are thus powerful and flexible tools for obtaining conditional cell distributions that preserve *expected* patterns in a target matrix, though they do not solve the problem of preserving exact marginals.

Exact preservation of higher-order properties is more difficult, and generally requires specialized algorithms. In the context of graph construction (viewing a binary adjacency matrix as a two-dimensional matrix of 0 or 1 counts), a large literature has emerged on methods for preserving row/column marginals (i.e., degree sequences), as well as degree mixing and block marginals (i.e., mixing rates); see e.g. [Tillman et al., 2019] for a review of several common cases. Construction algorithms - which produce an instance $\hat{N}$ exactly satisfying some target properties of N are of somewhat limited value for imputation, as they make no guarantees that the arrays constructed are representative of the set of feasible solutions (and generally they are not). Fortunately, however, it is often possible to construct Markov chain Monte Carlo (MCMC) algorithms that, given a feasible instance of $\hat{N}$, will simulate draws from a uniform (or other) target distribution over the set of feasible imputations. For our purposes, the most relevant work is that of Bunea and Besag [2000],

who provide an algorithm for sampling three-way count arrays that approximate a target distribution while preserving all two-way marginals. (When the two-way marginals are not available, Monte Carlo methods are available to construct data based only on one-way margins [Bunea and Besag, 2000], though we do not pursue this here.) We leverage and modify this procedure, using it to design an annealing algorithm that generates single imputations preserving both two-way marginals and higher-order correlations (a necessary goal for high-volume applications); in turn, we produce our target distributions using the log-linear modeling approach described above, exploiting the spatial hierarchy of areal unit data to obtain correlation information from higher-level units while preserving lower-level marginals.

The primary contribution of this paper is the implementation and development of a technique to impute three-way crosstab data that exactly preserves a set of integer marginals. Existing imputation techniques (including many of the ones discussed in this section), have difficulty with this kind of constraint structure. We leverage work on existing imputation techniques that allows for the incorporation of higher order spatial data to improve the quality of the imputed data.

3.3 Technical Description

3.3.1 Data Representation

As discussed above, we are interested here in the specific case of imputing an unknown three-way array of counts, $n \in \mathbb{N}^{I \times J \times K}$, for which the two-way marginals (i.e., quantities of the form $n_{ij\cdot}, n_{i\cdot k}, n_{\cdot jk}$) are known. This array is assumed to represent the cross-tabulation of entities within a given areal unit, for which the corresponding cross-tabulation of entities within a parent unit n^H, is fully observed. Our goal will then be to impute $n|\{n^H, n_{ij\cdot}, n_{i\cdot k}, n_{\cdot jk} : i \in 1, \ldots, I, j \in 1, \ldots, J, k \in 1, \ldots, K\}$, while satisfying all observed marginals.

3.3.2 Imputation Method

To impute the data contained in the three-way marginal array, we extend the work of Bunea and Besag [2000]. Using this algorithm as a baseline, we take a valid starting three-way array and use MCMC to simulate draws from the distribution of valid three-way arrays, given the set of two-way marginals that constrain it and a target distribution at a higher level of geography. We employ simulated annealing to both

find the valid starting point and to find a maximum-probability array with respect to the target distribution, a robust heuristic optimization procedure that helps avoid becoming trapped in local maxima. More details on the imputation process can be found in Algorithm 1.

The Target Distribution

Our algorithm, which is discussed in Section 3.3.2 requires a target distribution to be approximated (subject to our marginal constraints); since the two-way constraints will automatically account for all known information about the target array (n), the role of this distribution is to provide information regarding three-way associations that cannot be obtained for the target areal unit. We here employ the conditional log-linear model for the fully observed contingency table of the parent of the focal areal unit, n^H, to generate the target distribution. As a log-linear model is a discrete exponential family on the space of count arrays, it can be understood as leading to the maximum entropy distribution on the space of such arrays given the observed statistics and appropriate choice of reference measure [Darroch and Ratcliff, 1972, Jaynes, 1982]. Concretely, when applied to statistics based on table margins, it results in an inferred distribution that preserves the expected margins in the contingency table, while maximizing the uncertainty of the cell values given those expectations. Here, we base our target distribution on the three-way effects observed in n^H, while simulating conditional on the two-way margins of n; this gives a maximum-entropy approximation to the three-way structure of n^H, net of the (exactly preserved) marginal constraints of n, which allows us to use information from higher-order areal units to inform imputation for low-order units. This is accomplished as follows.

A saturated log-linear model contains sufficient statistics of effects at different levels in the contingency table. Specifically, for an array defined by three dimensions/covariates i, j, k, we can specify the model as:

$$\mathbf{E}(n_{ijk}) = \tau\tau_i\tau_j\tau_k\tau_{ij}\tau_{ik}\tau_{jk}\tau_{ijk}$$

where $\mathbf{E}(n_{ijk})$ denotes the expected count of the i, j, k cell; τ is the intercept, or the main effect of the contingency table; τ_i, τ_j, τ_k denote the marginal effects for the dimensions i, j, k, respectively; $\tau_{ij}, \tau_{ik}, \tau_{jk}$ denote the two-way interaction effects (with dimensions as above); and, finally, τ_{ijk} denotes the three-way interaction effect over all three dimensions. (Note that fixing the expectation, when combined with the assumption of a maximum entropy distribution over the set of possible matrices under a Poissonian reference

measure,[2] fully specifies the model.)

With information of two-way margins available for the target areal unit, one could estimate the marginal effects, and the two-way interaction effects. However, this is not sufficient to provide information about the three-way interaction term. Here, we approximate the three-way interaction effect for the contingency table of our target areal unit by the effect observed for its parent areal unit (treating the former *de facto* as a sample from the latter). This can also be viewed as a two-step process, where we first get an expected cell input based on information at the lower-level, and then re-calibrate it using information of the three-way interaction effect from the higher level. Formally, take $\mathbf{E}^L$ to be the expectation given all observable margins of the lower-level (i.e., target) areal unit; then we have

$$\mathbf{E}^L(n_{ijk}) = \tau^L \tau_i^L \tau_j^L \tau_k^L \tau_{ij}^L \tau_{ik}^L \tau_{jk}^L, \tag{3.1}$$

where $\tau_\cdot^L$ reflects parameter estimates based on the marginals of the observed (lower-level) areal unit. Now, letting τ_{ijk}^H be the estimate of the three-way effect from n^H, we employ the specification

$$\mathbf{E}(n_{ijk}) = \mathbf{E}^L(n_{ijk}) \tau_{ijk}^H. \tag{3.2}$$

Thus, we employ data from n^H to fill in the "missing piece" that cannot be obtained from n itself, while retaining all lower-order information from n.

Owing to the exponential family properties of the log-linear model, the parameters τ are easily estimated from the observed counts. The parameters describe the ratio between expected cell counts with and without the effects they represent; therefore, they are equal to 1 when absent. The general effect τ is equal to the grand mean of the contingency table, i.e. $\overline{n_{\cdots}}$. The one-way marginal effects are in turn equal to the ratios between the corresponding marginal means and the grand mean. Formally,

$$\tau_i = \frac{\overline{n_{i\cdot\cdot}}}{\overline{n_{\cdots}}}$$
$$\tau_j = \frac{\overline{n_{\cdot j\cdot}}}{\overline{n_{\cdots}}}$$
$$\tau_k = \frac{\overline{n_{\cdot\cdot k}}}{\overline{n_{\cdots}}}$$

where $\overline{n_{i\cdot\cdot}}$, $\overline{n_{\cdot j\cdot}}$, and $\overline{n_{\cdot\cdot k}}$ denote respective marginal means. The two-way interaction effects, in turn, are

[2] I.e., $h(x) = \prod_i (x_i!)^{-1}$, where the product is over cells. This amounts to assuming indistinguishability of individuals within groups.

equal to the ratios of the respective two-way means to the expectations of those means arising from the respective one-way means. Formally,

$$\tau_{ij} = \frac{\overline{n_{ij\cdot}}}{\frac{\overline{n_{i\cdot\cdot}}\,\overline{n_{\cdot j\cdot}}}{\overline{n_{\cdots}}}} = \frac{\overline{n_{ij\cdot}}\,\overline{n_{\cdots}}}{\overline{n_{i\cdot\cdot}}\,\overline{n_{\cdot j\cdot}}}$$
$$\tau_{ik} = \frac{\overline{n_{i\cdot k}}\,\overline{n_{\cdots}}}{\overline{n_{i\cdot\cdot}}\,\overline{n_{\cdot\cdot k}}}$$
$$\tau_{jk} = \frac{\overline{n_{\cdot jk}}\,\overline{n_{\cdots}}}{\overline{n_{\cdot j\cdot}}\,\overline{n_{\cdot\cdot k}}}$$

where $\overline{n_{ij\cdot}}$, $\overline{n_{i\cdot k}}$, $\overline{n_{\cdot jk}}$ denote the respective two-way means. Therefore, we may rewrite equation 3.1 in terms of observed counts as

$$\mathbf{E}^L(n_{ijk}) = \frac{\overline{n_{ij\cdot}}\,\overline{n_{i\cdot k}}\,\overline{n_{\cdot jk}}\,\overline{n_{\cdots}}}{\overline{n_{i\cdot\cdot}}\,\overline{n_{\cdot j\cdot}}\,\overline{n_{\cdot\cdot k}}}, \tag{3.3}$$

a quantity that is easily calculated.

With the first factor in hand, we now require only τ^H_{ijk}. As with the previous cases, the three-way interaction effect is equal to the ratio of the three-way marginal mean (here, identically the count of the i, j, k cell) to the expectation given the lower order effects.

Next, we process the three-way interaction effects using information from the higher-level unit. Similar to the previous derivations, the three-way interaction effect equals to the ratio of the cell with three-way interaction effect over that without the effect. Bearing in mind that all relevant counts here are for n^H, we have

$$\tau^H_{ijk} = \frac{\overline{n^H_{ijk}}}{\frac{\overline{n^H_{ij\cdot}}\,\overline{n^H_{i\cdot k}}\,\overline{n^H_{\cdot jk}}\,\overline{n^H_{\cdots}}}{\overline{n^H_{i\cdot\cdot}}\,\overline{n^H_{\cdot j\cdot}}\,\overline{n^H_{\cdot\cdot k}}}} = \frac{\overline{n^H_{i\cdot\cdot}}\,\overline{n^H_{\cdot j\cdot}}\,\overline{n^H_{\cdot\cdot k}}\,\overline{n^H_{ijk}}}{\overline{n_{ij\cdot}}\,\overline{n_{i\cdot k}}\,\overline{n_{\cdot jk}}\,\overline{n_{\cdots}}}, \tag{3.4}$$

which is again easily calculated from the observed arrays. In passing, we note that this expression for τ^H_{ijk} makes clear that it is already "normalized" with respect to the lower-order marginals of n^H; thus, differences between n and n^H in such quantities do not prevent τ^H_{ijk} from being used to model n (and, indeed, the three-way effects by construction do not affect any lower-order marginal expectations).

Putting these pieces together, the final target distribution is proportional to a product of Poisson distributions (a form that arises from the maximum entropy construction), whose expectations are functions of data from the target areal unit and its parent. The final target expectation for a given cell is the product of the

expected distribution given the lower level information (Eq. 3.3), and the three-way interaction effect from n^H (Eq. 3.4), i.e.

$$\mathbf{E}(n_{ijk}) = \frac{\overline{n_{ij.}}\,\overline{n_{i.k}}\,\overline{n_{.jk}}\,\overline{n_{...}}}{\overline{n_{i..}}\,\overline{n_{.j.}}\,\overline{n_{..k}}} \cdot \frac{\overline{n^H_{i..}}\,\overline{n^H_{.j.}}\,\overline{n^H_{..k}}\,\overline{n^H_{ijk}}}{\overline{n_{ij.}}\,\overline{n_{i.k}}\,\overline{n_{.jk}}\,\overline{n_{...}}}. \tag{3.5}$$

Imputating a Three-Way Array

While the target distribution that we specify above will be used to find a maximum probability array, any three-way array imputed must match observed two-way marginals. Thus, we separate our imputation procedure into two distinct steps. First, we construct an array that satisfies the constraints imposed by the two-way marginals. Then, with this array that matches observed marginals, we optimize the array with respect to the target distribution, preserving two-way marginals. Each of these components is non-trivial. Due to the integer constraints, finding an array that matches observed marginals (a valid array) is not possible using standard techniques (such as the expected count array formed when performing a generalized Chi-Square test). Likewise, for the three-way case, optimizing the array to maximize a target distribution is a challenging task.

Constructing a Valid Array

Our algorithm to impute a target three-way array begins by finding an array that satisfies the observed two-way marginals. This component must solve an array *construction* problem, prior to the optimization problem discussed in the second part of the algorithm (see section 3.3.2). This part of the algorithm is concerned only with satisfying the two-way and integer constraints, and does not consider the target distribution for array construction.

Our strategy (detailed with pseudocode in Algorithm 1) can be broadly described as follows. Algorithm 1 also includes optimizations discussed in Section 3.3.2. For a full description of the algorithm, see section 3.3.2. Our algorithm initializes an array using data from the zero-way marginal (i.e. the total array population). All population is divided equally across the array, with any remainder allocated to the first cell. This is detailed in lines 1 and 2 of Algorithm 1. This initial state ensures that the total population of the array and the integer constraints are satisfied. However, it is unlikely that this initial array state will satisfy the

constraints imposed by the one or two-way marginal values. We can define the deviation of our constructed array and the observed two-way marginals with the sum of the absolute values of the differences between the two-way marginals of our constructed array and the two-way marginals of the target array. We then seek to minimize this deviation.

We utilize a strategy of *simulated annealing* to produce a valid array from the initial state of our constructed array. Simulated annealing is a heuristic optimization technique designed to find the global minimum of an objective function, with minimal assumptions regarding the function and search space. This strategy will simulate moving values (individuals) between cells in the array, keeping track of the deviations between the simulated marginals values and the target marginal values. A single move will decrease the value of one cell and increase the value of another cell. However, the array is not considered as a valid state unless all cells in the array are non-negative. If there is a negative value in the array after a proposed move, we draw a new move based on the state of the proposed array. This process of drawing proposed changes will continue until a valid state of the array is drawn. This valid state will be proposed as a new state of our constructed array, and we compute the marginals of this array, as well as the arrays deviation from the observed marginals.

The annealer will always accept moves in the array that decrease the deviation between simulated and target marginals, as these moves will bring the simulated array closer to one that satisfies the constraints of the two-way marginals. The annealer will also accept moves that *increase* the deviation with a probability equal to $\exp(\frac{D_C - D_P}{T})$, where D_C is the deviation between marginals for the current array state, D_P is the deviation from the marginals for a proposed array state, and T is a temperature parameter. We still accept moves that increase deviation from the target marginals in order to prevent the annealer from finding a local minimum in the error space. However, the temperature parameter T will scale the likelihood that disadvantageous moves are taken. At high temperatures, accepting moves that increase our deviation is more likely, while lower temperatures make it much more difficult to accept these moves. The idea behind simulated annealing is to begin with a high temperature and allow the state of the array to vary more easily with respect to our deviation. This will help to prevent the state of our array from being stuck in a local minimum of deviation. As the annealer runs, we decrease the temperature geometrically, which will minimize the deviation by the end of the annealing run. Although convergence was easily obtained in the cases studied here, it should be noted that it is possible that the annealer will not converge to a a valid array. In this case, repeatedly restarting with a higher temperature and using a slower cooling schedule until convergence is obtained is a practical strategy. It should be noted that, regardless, convergence is always *verifiable*, since we can always determine whether or not our current array satisfies the target marginals (and, if not, the

degree of divergence).

While the annealer we describe here should produce a valid array that matches the constraints from the two-way marginals, the basic version of the algorithm requires us to recompute the two-way marginals every time we get a new state for our constructed array. While computing the marginals of the array once does not take a significant amount of time, computing them for every array state does add a high cost in computational time. To avoid this cost and improve the algorithm runtime, we introduce several optimizations, detailed in section 3.3.2.

Optimizations for the Construction Algorithm

The algorithm detailed in section 3.3.2 provides an array that will satisfy both the two-way marginal and integer constraints. However, due to the requirement to recompute the two-way marginals of proposed arrays, the algorithm can be expensive. To achieve better performance for larger datasets, we in practice implement a version of the algorithm that uses a change score. Specifically, we compute the *difference* between the initial state of the array and the target marginals, keeping track of these persistent errors. When we move values between array cells, we then update this persistent error by subtracting a person from the departure cell of the marginals, and adding a person to the arrival cell of the marginals (rather than recalculating the marginals anew). This persistent error is equivalent to using the error metric from section 3.3.2, but does not require recomputing marginals.

The updated error metric will avoid the computational cost of recalculating the marginals, but does require additional components. As noted above, we need to compute a map between the three-way array and each of the two-way marginals. This mapping will allow us to remove a person from the relevant cell of the two-way marginals and add them to the arrival cell when making a move in the three-way array. However, we only compute this mapping once, and can then refer to it when making moves in the three-way array. In Algorithm 1, the helper function mapIndexToMarginal will take a three-way array index and map it to the X,Y, or Z marginals respectively.

Description of Construction Algorithm

Algorithm 1 provides pseudocode for the construction of a three-way array that matches integer and two-way marginal constraints. This algorithm uses a set of two-way marginals X, Y, and Z. The name of the marginal

refers to the direction that we sum across the array to produce each marginal. We also need a set of helper functions for this algorithm. The functions xMargins(a), yMargins(a), and zMargins(a) each take a three-way array and produce a a two-way marginal. The function RandomInt(a,b) produces a random integer from a to b, inclusive. mapIndexToMarginal(a) takes a three-way array index and maps this index to a two-way marginal index. We use one of these functions for the X,Y, and Z marginals. Finally, numNegative(a) takes a three-way array and returns the number of negative values in the array.

Lines 1 and 2 of 1 produce the initial state of the array. The variable $numTotalCategories$ provides the number of cells in the array. Next, lines 3-5 produce the deviation of the initial state of the array from the observed marginals. These values will be used to compute an error metric, and will be used as a persistent deviation throughout the algorithm. Next, we specify how many arrays we will simulate with the annealer (M), and begin simulating arrays within the loop on line 6. Lines 7-10 are used to initialize the state of a proposed array, as well as the deviation that this proposed array would have from the target marginals.

After the deviation and array values have been initialized for our simulation, the second while loop (on line 11) begins the search for a valid array state to compare to our initial array state. Lines 12-16 draw two three-way array indices, i and j, and ensures that they are different. We also produce the corresponding two-way array indices for X,Y, and Z using lines 17-22, which use the mapIndexToMarginal() helper function, which uses a pre-computed map between three-way indices and two-way marginal indices. After we produce all of the necessary indices, we move a value in the three-way array from index i to index j, which simply adds one to the j^{th} cell of the array, and subtracts one from the i^{th} cell. Lines 25-30 also keep track of the move in the three-way array in the two-way marginals.

For this algorithm, a valid state of the array is one in which all cells are non-negative. At the end of every proposed move, we check to see if this condition is met (Lines 31-33), and if so, immediately end the search for a new move. If any array cells are negative, we draw another move and continue until we have a non-negative array.

With our non-negative array, we can next compute an error metric that uses the deviations of the initial array and the deviation of the proposed array. Line 35 computes this error, which is the difference in the absolute values of the deviation summed for the initial array and the proposed array. This value would be positive if the proposed array reduces the deviation from the target marginals, while it would be negative if the proposed array increases the deviation. The probability that our proposed array is accepted and becomes the current state of the array is computed on line 36, and is simply the error term divided by the

temperature term, exponentiated. Lines 37-42 check to see if we accept the proposed array state. If we do, then the proposed array becomes the current array. Likewise, the deviations from our proposed array state would become the initial deviations for the next iteration of the loop. If the array state is not accepted, the proposal is discarded and we begin from the initial state again.

Before the loop iteration ends, line 44 cools the temperature parameter. We use a geometric cooling schedule, where the temperature parameter is multiplied by a constant for each run of the array. The final state of the array is returned on line 46, and should have minimized the difference between the marginals of the array and the target array's marginals. In the implemented algorithm, we also check to see if the deviation between marginals is zero, and end the annealing process if it is. It is important to note that while this description has assumed that we are using this algorithm for single imputation, the algorithm will also work for multiple imputation. By fixing the cooling schedule and temperature parameters at 1, we are able to draw directly from the target distribution, which will be produced by the Markov chain.

MCMC Optimization Algorithm

The construction algorithm detailed above will produce a three-way array that satisfies the integer and two-way marginal constraints. The second component of our algorithm optimizes the three-way array's values with respect to a target distribution, using the specification described in Section 3.3.2. Our algorithm builds on the approach described by Bunea and Besag for simulating from three-way count arrays [Bunea and Besag, 2000]. This algorithm assumes that we start with a legal imputation of the target array (i.e. an array that satisfies the set of two-way marginals and has no negative values), as well as a target distribution.

Both our algorithm and the Bunea and Besag algorithm use MCMC to simulate three-way array configurations. The transitions between states use a *basic move*, in which a person is moved from one cell to another in the array. However, as each state of the array will is required to match the two-way marginal constraints, eight cells in the three-way array are modified in total for each basic move. Additionally, we can define the log-likelihood of a given array state under the target distribution, which we will denote $l(n)$, where n is a three-way array. For two arrays, a current and proposed array, the probability for the Markov chain to accept the proposed move is $\exp(l(n') - l(n))$, where n is the current array state, and n' is the proposed array state. Because the basic move preserves all lower-order marginals, validity of an array produced by this method is guaranteed (i.e., any array generated from a valid starting point will always be a valid array).

Algorithm 1 Produce a three-way array that satisfies a set of two way marginals X, Y, Z

$n_{ijk} \leftarrow$ floor(sum(X) / $numTotalCategories$)
$n[1] \leftarrow n[1]$+sum(X)−floor(sum(X)/$numTotalCategories$)
$xError \leftarrow$xMargin(n) − X
$yError \leftarrow$yMargin(n) − Y
$zError \leftarrow$zMargin(n) − Z
while $M > 0$ **do**
  $n' \leftarrow n$
  $xError' \leftarrow xError$
  $yError' \leftarrow yError$
  $zError' \leftarrow zError$
  #Ensure that the next state of the array has no negative values
  while $K > 0$ **do**
    $i \leftarrow$RandomInt(1, $numTotalCategories$)
    $j \leftarrow$RandomInt(1, $numTotalCategories$)
    while $i == j$ **do**
      $j \leftarrow$RandomInt(1, $numTotalCategories$)
    end while
    # Map the origin and destination of the move to the marginals
    $mappedI_X \leftarrow$mapIndexToMarginalX(i)
    $mappedI_Y \leftarrow$mapIndexToMarginalY(i)
    $mappedI_Z \leftarrow$mapIndexToMarginalZ(i)
    $mappedJ_X \leftarrow$mapIndexToMarginalX(j)
    $mappedJ_Y \leftarrow$mapIndexToMarginalY(j)
    $mappedJ_Z \leftarrow$mapIndexToMarginalZ(j)
    #Do the move in the three-way array
    $n' \leftarrow$moveAPerson(n, i, j)
    $K- = 1$
    # Update the marginal deviations
    $xError'[mappedI_X] \leftarrow xError'[mappedI_X] - 1$
    $yError'[mappedI_Y] \leftarrow yError'[mappedI_Y] - 1$
    $zError'[mappedI_Z] \leftarrow zError'[mappedI_Z] - 1$
    $xError'[mappedJ_X] \leftarrow xError'[mappedJ_X] + 1$
    $yError'[mappedJ_Y] \leftarrow yError'[mappedJ_Y] + 1$
    $zError'[mappedJ_Z] \leftarrow zError'[mappedJ_Z] + 1$
    #If our array is non-negative, end the search for a move
    if $numNegative(n') == 0$ **then**
      break
    end if
  end while
  #Evaluate the relative error of our proposal and current arrays
  $error \leftarrow$sum(abs($xError$),abs($yError$),abs($zError$))−sum(abs($xError'$),abs($yError'$),abs($zError'$))
  $transitionProbability \leftarrow \exp(error/temperature)$
  if uniform(0, 1) < $transitionProbability$ **then**
    $n \leftarrow n'$
    $xError \leftarrow xError'$
    $yError \leftarrow yError'$
    $zError \leftarrow zError'$
  end if
  $M- = 1$
  #Cool the chain
  $temperature \leftarrow temperature * 0.9$
end while
$return(n)$

We modify Bunea and Besag's algorithm by using simulated annealing. In the construction component of this algorithm, we used annealing to optimize the state of an array to minimize deviation from the observed marginals. As our goal in this part of the algorithm is to find the most likely configuration of the algorithm under the target distribution, we can use simulated annealing for better single imputation quality. The addition of simulated annealing is straightforward, and only requires us to modify the acceptance probability with a temperature parameter. As with the construction algorithm, the temperature parameter will scale the probability that the optimization algorithm accepts proposed arrays that are less likely under the target distribution. Higher temperature values (above 1) will increase the likelihood that less likely array configurations are accepted. Likewise, as temperatures approach zero, the probability of accepting lower-probability array configurations also goes to zero. A fixed temperature at 1 will accept new array states by exactly evaluating the likelihood within the target distribution. A benefit of updating this algorithm to use simulated annealing is that the algorithm can be run in both single and multiple imputation modes. In single imputation mode, the temperature would be set above 1, and the cooling schedule would be set below 1. However, as mentioned above, by setting the temperature and cooling schedule to 1, we would draw directly from the target distribution, which enables multiple imputation.

Optimization Algorithm Description

Algorithm 2 provides pseudocode for the optimization of a three-way array with respect to a target distribution corresponding to the maximum entropy distribution on n conditional on $\mathbf{E}(n)$ and the two-way marginals. We need several helper functions for this implementation. First, we use a helper function doBasicMove(n), which takes a three-way array as an input, and moves someone from one cell in the array to another, maintaining all two way marginals. As discussed above, this basic move modifies eight cells of the three-way array, and is described in more detail in Bunea and Besag [2000]. We also use the helper function numNegative(n), which takes a three-way array as an input, and outputs the number of negative values in the array. Finally, the helper function numNegativeOne(n) takes a three-way array as an input, and oututs the number of negative ones in that array.

The implementation of our algorithm relies on a Markov chain that cools as the annealer runs. The total length of the Markov chain is $M * L$, where M is the number of times we cool the Markov chain, and L is the number of iterations we run the Markov chain at each temperature. At each step of the Markov Chain, we start by proposing a basic move on the target three-way array n (Line 3). n', the proposed three-way array, will match all two-way marginal and integer constraints. Next, we follow the Bunea and Besag algorithm by

checking the number of negative values (specifically negative ones) in n'. If there is exactly one negative one in the array, we draw a new basic move from n', and continue to do so until either there is either more than one negative value in n', or n' becomes a non-negative array. If n' is an array with more than one negative value, we discard the proposal, keeping the original state of the array. However, if n' is a non-negative array, we compute the ratio of the likelihoods for n' and n (in log space) under the target distribution, and divide this log-likelihood by a temperature parameter (Line 11). When exponentiated, this is the probability that the proposed array is accepted as the next state of the Markov chain.

Every L iterations of our Markov chain, we cool the chain. Like the construction algorithm, we use a geometric cooling schedule for this algorithm, multiplying the temperature of the annealer by a constant every L iterations. As the temperature of the annealer decreases, the Markov chain will accept proposed states that are lower likelihood under the target distribution less often than at higher temperatures.

Algorithm 2 Impute a three-way crosstab

Require: target expectations $\mathbf{E}(n)$, array state n, initial temperature T, decay parameter $c = 0.94$

```
1:  while M > 0 do
2:    while L > 0 do
3:      n' ←doBasicMove(n)
4:      while numNegativeOne(n') = 1 do
5:        n' ←doBasicMove(n')
6:      end while
7:      if numNegative(n') > 1 then
8:        next
9:      else {numNegative(n') = 0}
10:       randomNum ←Uniform(0, 1)
          #Accept the proposed array with probability equal to the ratio of probabilities of proposed:current
          arrays
11:       if randomNum <min(1, exp((l(n')−l(n))/T))) then
12:         n ← n'
13:       end if
14:     end if
15:     L− = 1
16:   end while
17:   M− = 1
      #Cool the chain
18:   T ← T ∗ c
19: end while
20: return(n)
```

3.4 Validation of Imputed Data Quality

Above, we have provided algorithms for construction and imputation of three-way count arrays with targeted characteristics. Next, we describe the test imputations and metrics that we use to validate the quality of population data imputed using this approach. Our validation tests employ U.S. census data on population distributions, using several levels of geographic aggregation. We also use two validation metrics to determine data quality.

3.4.1 Data used for Validation Runs

We use data from the 2010 U.S. census to assess the quality of our imputation technique. The U.S. Decennial Census published complete three-way population distributions at several levels of geographic aggregation. The Census uses a geographic hierarchy for their data products, with Census blocks aggregating into Census tracts, which themselves aggregate into counties.[3] We consider the three-way distribution of race, gender, and ethnicity within each geographically defined subpopulation (i.e. count data for each three-way category). Ethnicity has two categories, Non-Hispanic and Hispanic. Gender also uses two categories, Male and Female, while Race has 7 categories[4]. Given the national scale of the Census, these distributions provide a large dataset for us to test our imputation.

We perform two imputation studies in order to validate the approach. The first employs U.S. Census data across the entire United States, specified at the county and tract level. Here, target distributions are defined using three-way distributions at the county level. At the tract level, we use the two-way observed marginals for our target array. Full three-way distributions are publicly available at the tract level, which allows us to directly validate the quality of the imputation. We impute the three-way distribution of race, ethnicity, and gender for each of the 73,057 tracts in the United States. We perform both single imputation and multiple imputation for each tract, comparing true values against imputed counts.

Our second imputation study involves analysis of a social outcome (exposure to crime), using data specified at the tract and block levels. We specify a target distribution using full, three-way arrays available at the tract level, and use marginal two-way arrays that were published at the Census block level of aggregation. We chose to only impute the three-way arrays for one U.S. state (California), which contains 710,154 census

[3]There is also an intermediate level of aggregation known as the block group, but since their data availability is similar to blocks, we do not consider them here.

[4]Those categories being White, Black, Asian, Native American, Pacific Islander, Multiple Races, and Other

blocks. Comparison of analysis at the tract level on actual versus imputed data provides another check on imputation quality. Although we cannot directly validate block-level imputation (since the three-way marginals are no available), we employ this for an illustrative case study described in section 3.4.3.

3.4.2 Imputation Parameters

For both of the validation samples, we use the same settings for both array construction and optimization. Additionally, the imputation calculations were completed on the same machine, using the same computing resources (facilitating timing comparisons).

First, we detail the parameters used for construction of a valid array. For each array we construct, we simulate up to 1,000,000 array states (M in Algorithm 1). We also allow for up to 1000 moves to find a new valid array state (K in Algorithm 1). We initialize our temperature parameter T to be ten times the error rate (deviation from the marginals) produced by the initial state of the array. In practice, all arrays produced by the implementation of Algorithm 1 were found to match the two-way and integer constraints, indicating that these parameters are sufficient for the heuristic optimization to succeed in finding valid array states.

Next, in the optimization component of the algorithm, we use a Markov chain of length 50,000 for each array. We cool this chain every 1,000 iterations, for a total of 50 annealing steps. The initial temperature parameter T is set to 10, with a cooling parameter c of 0.94. This allows the annealer to accept less likely array states more readily for half of the Markov chain, with the second half of the Markov chain behaving more strictly as a hill climber, seeking the maximum likelihood array state.

When doing multiple imputation, we fix both the temperature and the cooling parameter at 1 (i.e., we fix the algorithm at the target distribution, with no cooling). This allows us to draw directly from the distribution of array states that is specified by the target distribution and the marginal constraints. We use a thinning parameter of 1000 and a burn-in parameter of 1000, which were found to be adequate for convergence. In seeding the Markov chains for the multiple imputation draws, we initialize the optimization portion of our method with a draw from the single imputation mode of the algorithm. We do this to ensure that the Markov chain will burn-in, by starting it at a mode of the target distribution.

Each of our imputation studies was performed on an Intel Xeon E5-2599 V4 CPU. As the three-way array that is present for any areal unit does not depend on any other areal unit, this problem is trivially parallelizable.

Thus, we used 30 cores for each imputation. We also introduce several special cases where we can directly solve the state of the three-way array. The first case is the one where there is no population in the array. Second, we can directly solve the "one-hot" case, or arrays in which there is only one cell with any population. We can solve these arrays using only information from the two-way marginals, trivially imputing the array.

3.4.3 Metrics for Assessing Data Quality

We use two main methods for the assessment of data quality. Both of these metrics require observed data as a baseline. Therefore, we rely on the tract level imputation that is described above, as the full three-way tract arrays are published in addition to the two-way marginal data. Our first metric for assessing data quality relies on an error metric, and can be assessed on an individual array basis. We also provide a metric for assessing quality that depends on stable performance in a downstream analysis.

Error-Based Metric for Data Quality

The first metric we use to asses data quality measures the degree that an imputed three-way array departs from its observed values. In other words, this accuracy metric measures how many people are mismatched between a simulated and observed array. Our error metric, E is defined as $E = \sum_i |O_i - I_i|$, where O is the observed array for an areal unit, I is the imputed array for the same areal unit, and the sum is over entries of the array. This error metric simply represents the number of people who are misallocated by the imputation.

For purposes of expressing this error metric in a standardized manner, we divide E by the number of people present in the areal unit, which normalizes the error values to a range between 0 and 1. (In tracts where there is zero population, we define the metric to be zero.) This value is referred to as E_R, and describes the percentage of the tract that has been misallocated. Low values of this error metric indicate high quality imputed data.

We compare the errors produced by our algorithm to error rates produced by several other approaches. The first alternative to which we compare is the one described by Bunea and Besag, in which there is no simulated annealing, and we simply use a Metropolis algorithm to take a single draw from the distribution defined at the higher level of geography. We would expect that in this case, error rates would be broadly similar - since, if the Markov chain is burned in correctly, a random draw from that distribution is relatively

likely to be from a high-probability region - but with higher excursions due to the fact that the algorithm will occasionally select plausible but low-probability arrays. This provides a point of comparison for the annealing algorithm, which uses the same target distribution but attempts provide a maximum-probability array.

Next, we examine the case where we use the expected values provided by the target distribution as the final imputed values. These expected counts are produced using the log-linear framework described in section 3.3.2, which incorporates two way data from the target level of geography with three-way patterns present at the higher level of geography. Because the log-linear model is not constrained to satisfy integer constraints, it is expected to accumulate numerous errors; however, it nevertheless incorporates distributional information, and (being easy to compute) is an obvious practical alternative.

Finally, to examine the improvement produced by simulating the distribution of possible three-way states under the target distribution, we also examine the error rates when using the array generated by the construction algorithm as the final imputed value. This array will be "valid," in that it satisfies both integer constraints and the known two-way marginals, but not otherwise adjusted. We specifically examine this case to better understand how the space of three-way arrays may be constrained by the two-way marginals. This technique would also omit all data from the higher level of geography, so we can examine how only incorporating the local demographic effects (i.e. the two-way marginal constraints) may produce different arrays from the observed data.

Case Study for Quality Checking

While direct error assessment is the most natural way to evaluate imputation quality, it does not speak directly to downstream impacts on subsequent analysis: relatively poor imputation may in some cases prove adequate when downstream analyses are robust, while sensitive analyses may require very high degrees of imputation accuracy. While such sensitivity inevitably depends on the analysis involved, we here use a case study involving a spatially heterogeneous outcome - exposure to crime in one's vicinity - as a plausible example of how errors may or may not impact substantive conclusions. Specifically, we carry out our analysis at the tract level using both observed and imputed data, allowing us to compare results obtained in the two cases. For this purpose, we employ both single and multiple imputation, allowing us to compare the performance of both estimators at recovering observed-data results. Finally, as an illustrative procedure, we repeat our data analysis at the block level. Although not suitable for validation (since we do not have

block-level observations), this analysis provides an example of how the imputation approach might be used in a realistic case, and how pushing analysis to finer levels of geographic detail can potentially impact our substantive conclusions.

Our case study examines how exposure to crime near one's home is related to one's demographic characteristics. Crime is heterogeneously distributed, making members of some groups more likely than others to be exposed; such exposure may, in turn, feed concerns about neighborhood safety, willingness to access local affordances, and stress. To examine this association, we use crime data obtained from police agencies for the Southern California Crime Study (SCCS). In the SCCS, the researchers made an effort to contact each police agency in the Southern California region and request address-level incident crime data for six part 1 Uniform Crime Report (UCR) categories: homicide, aggravated assault, robbery, burglary, motor vehicle theft, and larceny. These data come from crime reports officially coded and reported by the police departments and provide locations of crime incidents around 2010 covering about 83% of the population in a five-county area (Los Angeles, Orange, Riverside, San Bernardino, and San Diego). Crime events were geocoded for each city separately to latitude/longitude point locations using ArcGIS 10.2, and subsequently aggregated to various units such as blocks and tracts. The average geocoding match rate was 97.2% across the cities, with the lowest value at 91.4%. These data have been used in several prior studies [Kubrin and Hipp, 2016, Kubrin et al., 2018]. We specifically use the number of violent crime events that take place (homicide, aggravated assault, robbery), and compute the average over three years (2009-11) to smooth year to year fluctuations. Prior literature shows that one's exposure to crime is affected by many demographic features, including the ones that we have imputed in this paper [Alba et al., 1994, Logan and Stults, 1999, McNulty, 1999]. The actual form of this relationship has not been particularly closely examined, however - particularly at the level of small areal units, which are needed to avoid averaging across areas with different crime rates. Thus, using both observed and imputed data at the census tract level (see section 7.1 for details on the imputation), we specify a saturated linear regression model (i.e. main effects, two-way interaction terms, and three-way interaction terms). Additionally, we specify the same models at the census block level, only using imputed data. We examine the block level model to compare whether the effects are similar to those of the tract-level model. At the tract level the mean number of crime events in the data is 42.38 events, with a minimum of zero events and a maximum of 666 events. At the block level, we use an additional buffer around each areal unit. This buffer has a radius of 1km. The mean number of crime events for the block level data is 112.22 events, with a minimum of zero events and a maximum of 2234 events.

The census geographies that we use here are adjacent levels of the census spatial hierarchy. Census tracts

compose counties, and are often relatively large. For tracts represented in the SCCS, the average tract population in 2010 is 4604 people. While tracts can provide an overview of population distributions across space, the census block level is much more granular (and is often about the size of a city block). The average population for a census block in the area represented by the SCCS is approximately 80 people, while the five counties have an average population of 3.765 million.

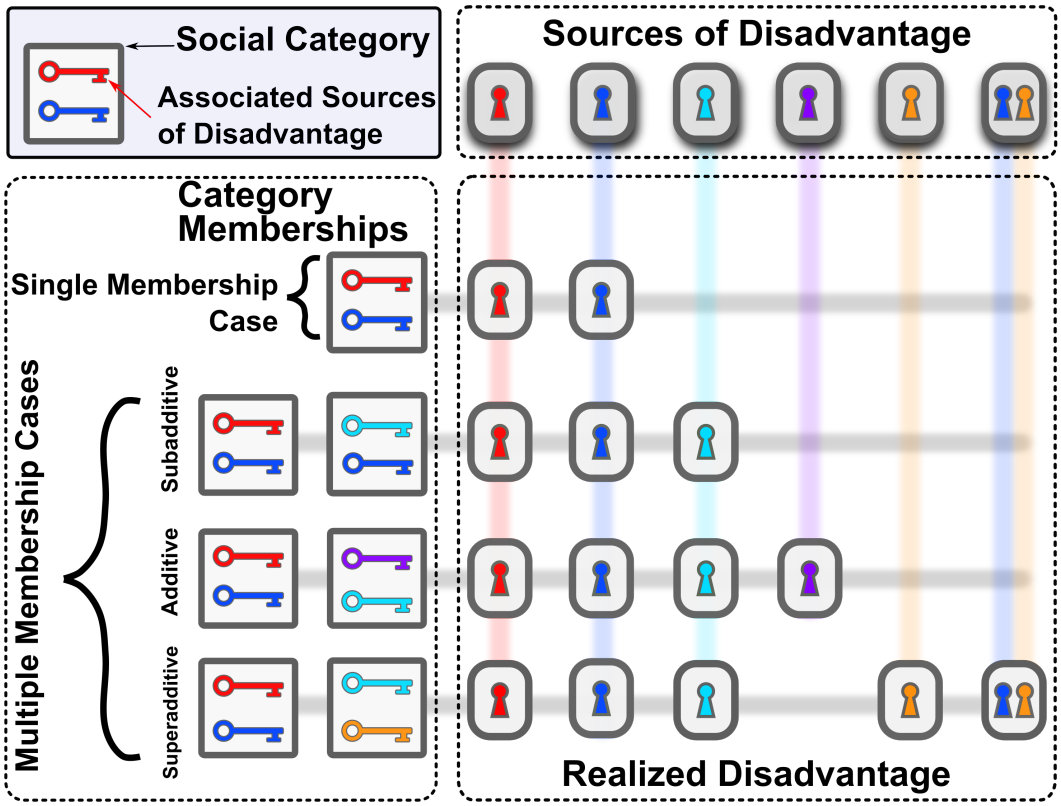


Figure 3.2: Schematic depiction of the ways in which overlapping social category memberships can lead to different degrees of realized disadvantage. Each social category (i.e. a race/gender/ethnicity category) has a set of associated sources of disadvantage. These sources of disadvantage can combine in a variety of ways. In the subadditive case, overlapping sources of disadvantage only contribute once to the total degree of disadvantage. In the additive case, all sources of disadvantage contribute once to the total amount of realized disadvantage. Under superadditivity, additional disadvantage is "unlocked" due to having multiple sources of disadvantage in distinct social categories.

In order to compute one's exposure to crime (i.e. the response term for this model), we use the number of crime events that occur in a given areal unit, using data from the SCCS data source. From this number of crime events, we consider a group's exposure to crime in that areal unit as $C_{ijk} = Ep_{ijk}$, where E is the number of violent crime events that occur for that areal unit, and p_{ijk} is the proportion of the total number of people in a three-way race, ethnicity, gender category (for the entire sample) that are present in the areal unit. These values are summed across all areal units to produce an exposure for each of the groups. This exposure functionally behaves as a weighted average of crime exposures in each areal unit. We then predict

this exposure using dummy variables for each race, ethnicity and gender category, as well as all two and three way interaction effects using these terms. The White, Non-Hispanic Male category is used as a reference group for these regressions.

Given that these fully specified models are not common in the literature, we present three hypotheses about the nature of the relationship between the explanatory factors and one's exposure to crime, motivated by more general notions of cumulative, intersectional, and saturated mechanisms of disadvantage. These hypotheses are schematically represented by Figure 3.2. We consider individuals as belonging to one or more *social categories,* reflecting e.g., race, gender, etc. Members of a given social category may be, on average, particularly likely to be exposed to specific *underlying sources of disadvantage*; some such sources may be unique to specific categories, while others may be shared by members of multiple categories. Schematically, Figure 3.2 depicts social categories as boxes, each of which contains a set of "keys" that "unlocks" particular sources of disadvantage (here, indicated by color). An individual belonging to only one social category receives the keys - and hence the sources of disadvantage - for that category. Where an individual belongs to multiple categories, they inherit the keys from each category they belong. The consequences of this can vary, leading to several hypothetical scenarios.

Our first scenario, represented by the *subadditive* row of Figure 3.2 involves the case where the sources of disadvantage for an individual's social categories overlap. In this case, having multiple memberships in disadvantaged groups provides more disadvantage than being a member of a single category, but not as much as the independent combination of both groups. Here, the sources of disadvantage *saturate*, and their effect on the individual is subadditive.

Our second scenario, represented by the *additive* row, occurs when there is no overlap in the sources of disadvantage for the categories to which an individual belongs. Here, the total disadvantage is simply the sum of the disadvantage for each category.

Finally, in our third scenario (the *superadditive* row) we consider the possibility that there are sources of disadvantage that require "keys" from multiple categories to unlock. In this case, the total disadvantage for multiple group memberships can exceed the sum of the group disadvantages, since a joint member is impacted by both the union of the two group sources and additional sources of disadvantage that arise from co-membership. This is often discussed in the literature within the context of *intersectionality* [Crenshaw, 1990], with the notion that belonging to multiple disadvantaged groups can have a substantially greater impact than the independent effects of each membership alone.

In the context of exposure to crime, it is plausible that sources of disadvantage associated with gender, race, and ethnicity could correspond to any of these three scenarios. To quantify this, we specify an *additivity index*, which we use to categorize the relationship for each of our three-way categories included in the model. This index can be defined by:

$$A = \frac{\beta_{ijk}}{a+\beta_i+\beta_j+\beta_k+\beta_{ik}+\beta_{jk}+\beta_{ij}}$$

where a is the intercept term, and the β terms are the regression coefficients for the one, two and three-way effects. If the three-way effect in this term is zero, then the index will also be zero, which implies a purely additive relationship. Likewise, if the sign of the total two-way effects and three-way interaction term are the same, then the index will be positive, indicating a superadditive relationship. Finally, if the numerator and denominator are of different signs, this index would be negative, which indicates a subadditive relationship. In the rare case where the total two-way effect (denominator) is zero, we define A to be zero.

The magnitude of the index is also informative. Usually, we would expect A to be between -1 and 1, which indicates that the three-way effect is smaller in magnitude to the rest of the effects. However, in the event that A is greater in magnitude than -1 or 1, this indicates that the three-way effect outstrips the combined two-way effects, and would be able to flip the sign of the total effect.

3.5 Tract Imputation Results

Next, we describe the results of the imputations discussed above, evaluating the overall quality of the imputed arrays. Imputing all 73,057 tracts took 7 hours and 33 minutes and 49 seconds on 30 cores of an Intel Xeon E5-2599 V4 CPU. As the tract-level three-way arrays in the United States are known, we can directly compare the imputed three-way arrays with to the observed data. We use the error metric specified in Section 3.4.3.

Array approximation results. This error metric provides support for the quality of the data imputation. Figure 3.3 describes the distribution of relative errors, showing that most tracts have a very low error rate. Given that the mean relative error is 0.8% and the 97.5th percentile of the error is 2.5%, this imputation schema produces three-way arrays that are excellent proxies for the observed data (with error rates at or below error rates in the Census itself [Khubba et al., 2022]).

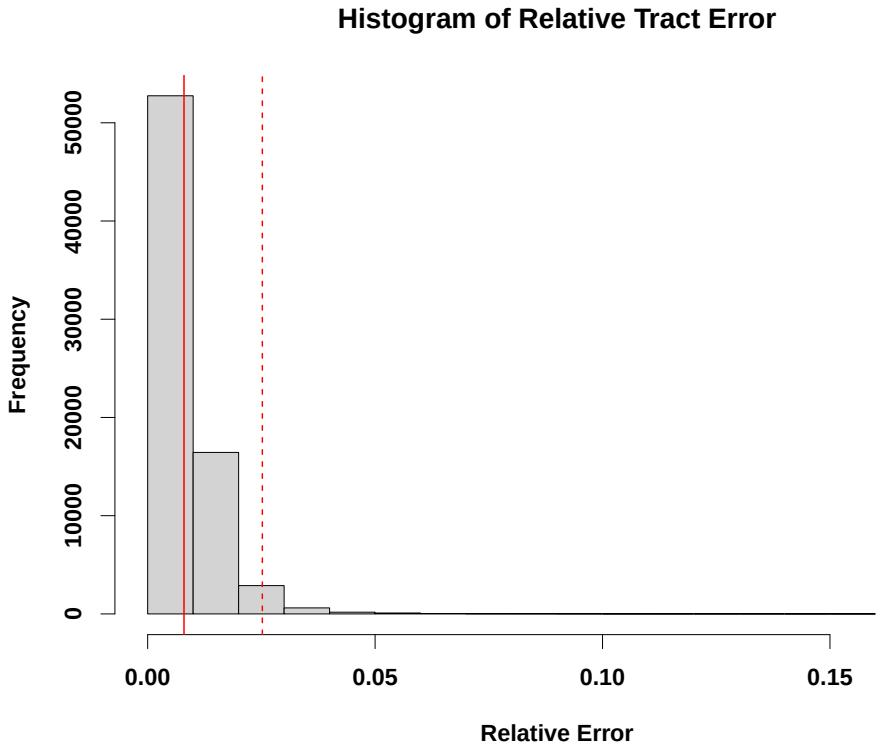


Figure 3.3: A histogram of relative errors. The solid red line is the mean (0.8%), while the dashed red line is the 97.5^{th} percentile (2.52%)

We compare these error rates to the rates obtained by the other procedures described in section 3.4.3. We find that the case where we draw directly from the target distribution (using Bunea and Besag's algorithm without simulated annealing) produces very slightly elevated error rates to the ones produced by our updated algorithm. The mean error rate for all tracts in the U.S. is 0.009 (0.9%), while the 97.5th percentile of the error is 0.0283 (2.83%). The arrays produced by this algorithm are produced by the same process that we use for multiple imputation, which we will also show produces similar qualitative results to the single imputation case when doing downstream data analysis. We thus conclude that there is some gain from annealing to find the mode of the target distribution (versus using an arbitrary draw), but error rates are not very sensitive to this aspect of the algorithm.

Next, when using the expected counts produced by the loglinear models (for our target distribution) as the imputed arrays, we see noticeably elevated error rates. The mean error rate is 0.0124 (1.24%), while the 97.5th percentile of this error distribution is 0.0396 (3.96%). While these error rates are still relatively low, they are roughly 50% higher than the annealed imputation method, and the estimates do not satisfy integer constraints (making them unsuitable for some applications).

For the third case, where we simply construct a valid three-way array that conforms to the integer and marginal constraints, we would expect the error rates to be significantly higher than for the case where we use simulated annealing to produce the most likely three-way array under the target distribution. Indeed, the mean error rate produced by this imputation is 0.047 (4.7%), while the 97.5th percentile of the error is 0.158 (15.8%). This case provides an interesting point of comparison, as it shows that the space of three-way arrays is significantly constrained by the two-way marginals, but despite this, there are still significant improvements that are made through the optimization components of the algorithm.

Overall, these results suggest that while constraints are powerful, incorporating distributional target information is still important for getting high-quality approximations. Given that this is done, optimization to ensure that a mode is selected (versus a random draw from the target) is helpful, but less vital. This also implies that our approach is not extremely sensitive to annealing performance, which may be useful in settings for which the cost of high-quality annealing runs is a concern.

Results for downstream analysis. In addition to direct approximation error, we also use the case study described in Section 3.4.3 to evaluate data quality. Our case study examines the effects of disadvantaged social category membership on exposure to crime. As we are using three-way arrays to examine this relationship, we are particularly interested in the three-way coefficients from the regression specified above. The coefficients for both the observed data model and the imputed data model are visualized in Figure 3.4. For the observed data model, the means and variances are directly computed from 1000 bootstrapped samples of areal units. We use a standard bootstrap sampling design for the observed data model. For the imputed data model, we examine the effect of using the algorithm in single imputation mode vs. multiple imputation mode. In single imputation mode, we draw a single array for each areal unit. Then, we sample areal units using the standard bootstrap design. In the multiple imputation mode, we use a slightly different sampling method. For each of the 2000 bootstrapped samples, we draw a set of arrays from the distribution defined by our target distribution. The Markov chains used to draw from this distribution were seeded with a draw from a single imputation run of the algorithm to ensure that the chains were burned in adequately. Then, we draw n arrays from that areal unit, where n is the number of times that areal unit has been drawn for that bootstrap sample. We then use the quantile method to compute the distribution of each coefficient.

The three-way coefficients from Figure 3.4 almost all overlap with zero, indicating mostly *additive effects*. In addition, the simulated and observed distributions of three-way coefficients all have significant overlap with each other. Further, in interpreting the effects, a researcher would obtain similar qualitative results

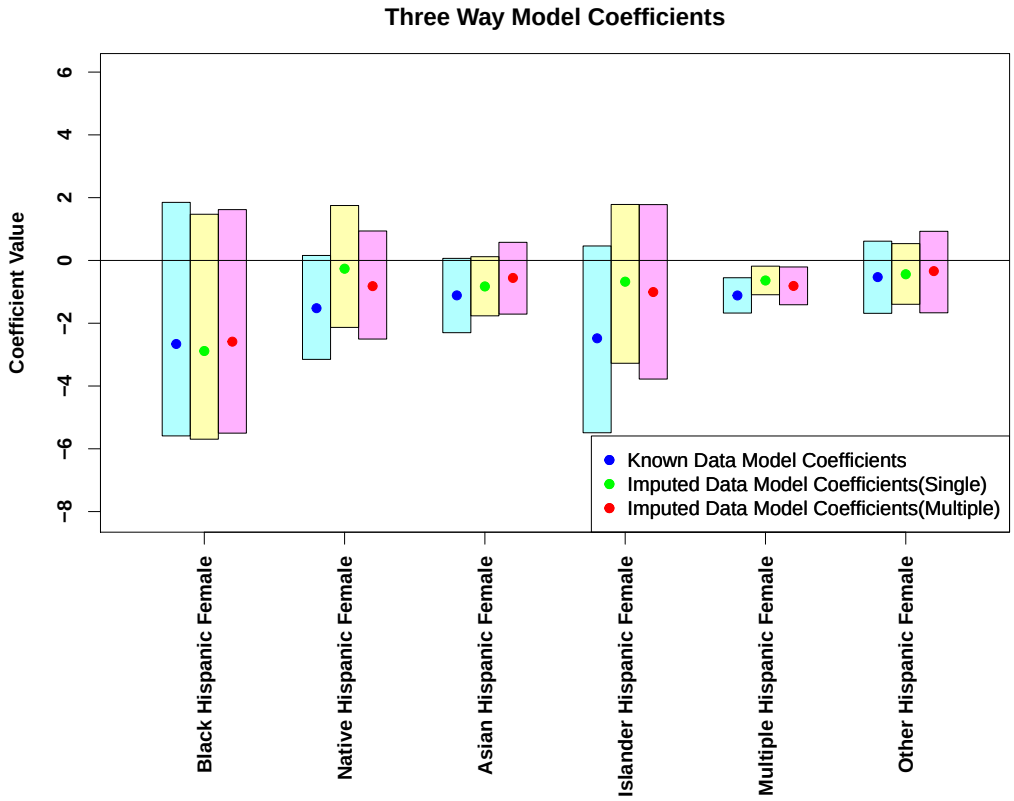


Figure 3.4: A plot of three-way effects where the blue points are the coefficients of the known model with cyan 95% simulation intervals, and the red points are coefficients of the multiply imputed model with magenta 95% simulation intervals. The green points and yellow 95% simulation intervals are for a model that uses a bootstrap design, but with single imputation rather than multiple imputation. Known data simulation intervals were computed with 2000 bootstrap iterations using the quantile method. Imputed model intervals (red) were computed using a set of MCMC samples that utilize the multiple imputation mode of the algorithm, while the yellow intervals use a single imputation mode of the algorithm.

from interpreting the observed and simulated models (i.e. the effects significantly different from zero are the same). This indicates that qualitatively, the simulated arrays are similar in nature to the observed arrays, and would not introduce significant error into downstream analysis. Additionally, while the single imputation produces coefficients that tend to be slightly closer to zero, both single and multiple imputation modes produce similar results to the observed data model.

We also examine the patterns in the coefficients reported in Figure 3.4, with respect to their additivity indices. For each of the coefficients reported from the analysis done with the bootstrapped samples from the observed three-way arrays, we compute the additivity index from Section 3.4.3. The distribution of these additivity indices is reported in Figure 3.5. As expected from a cursory examination of the coefficients in Figure 3.4, the majority of the coefficients exhibit an additive pattern. For the Black Hispanic Female and Pacific Islander Hispanic Female coefficients, there are some cases of subadditivity present, but for the most

part the three-way coefficients describe an additive relationship between disadvantaged category membership and exposure to crime at the tract level. Notably, we find no sign of systematic superadditivity at this level of aggregation.

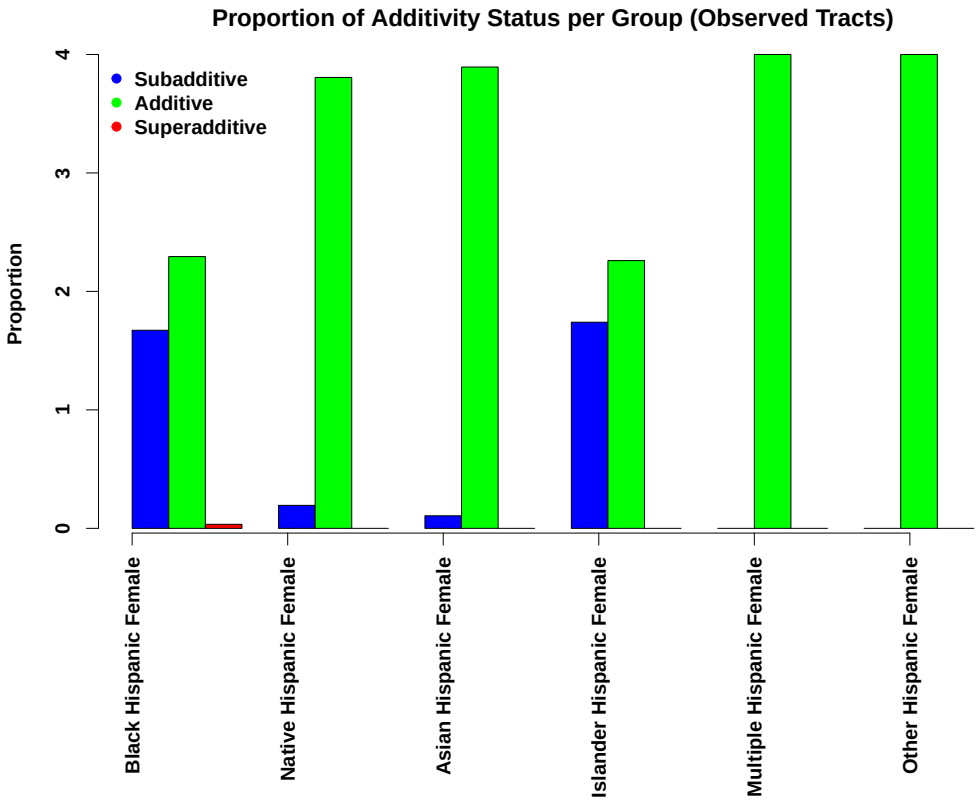


Figure 3.5: Additivity indices for each of the three-way categories in the model across 2000 bootstrap iterations. Additive values were ones in which the three-way effect size was less than 5% of the combined one and two-way effects.

3.5.1 Block Level Imputation Results

While the tract-level analysis suggests additive effects to be predominant, it is possible that this is an artifact of aggregating over locally heterogeneous units. Although the full data needed to replicate the observed-data analysis at the block level is not available, we can do so using our imputation scheme. For a single imputation, we impute the 710,145 census blocks in California on 30 cores of an Intel Xeon E5-2599 V4 CPU in 12 hours, 53 minutes, and 42 seconds. We hypothesize that the higher areal unit/second imputation rate is likely due to the lower population of census blocks in comparison with census tracts. There are also more census blocks that can be trivially solved (see Section 3.4.2) than census tracts. The full three-way arrays at the census block level are not available, so we are unable to compute the same error metrics that we use

for the census block imputation.

Given the imputed block-level arrays, we once again examine the relationship between gender, ethnicity, and race on exposure to crime. We apply the same basic procedure as in the tract level case study. However, rather than measuring crime in the specific block (which is too small a unit of analysis to provide a reasonable notion of exposure), we measure crime in a 1km buffer around each focal block. Once again, we use multiple imputation to get a set of potential arrays for each areal unit, drawing from the target distribution specified at the tract level.

The three-way effects from the block level exposure to crime are summarized in figure 3.6. We used 500 bootstrapped samples from the crime and areal unit data to compute the simulation intervals for this plot. The patterns in these coefficients generally match the patterns from the tract level analysis, but the magnitude of these coefficients is much greater. None of the three-way coefficients intervals contain zero, representing significant three-way effects. We use the same multiple imputation sampling that is described above to generate these estimates.

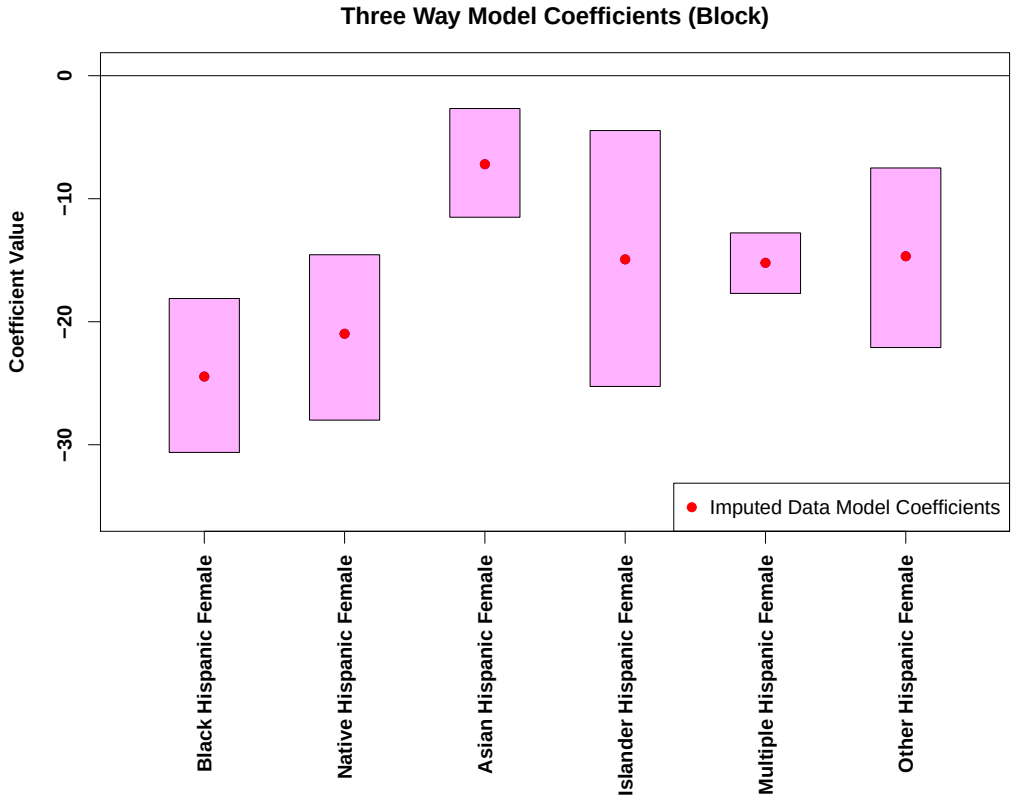


Figure 3.6: Estimates of three-way coefficients at the census block level. The red points are the mean estimates across 500 bootstrap intervals, with the magenta region representing the 95% simulation intervals.

As as the observed effects are consistently negative, we expect that there is less of a strongly additive pattern at the block level. Figure 3.7 depicts the patterns in the additivity index for the three-way coefficients. Almost all of the coefficients exhibit a strongly subadditive pattern, with only the Other race Hispanic Female and Asian Hispanic Female categories having some additive indices. We thus find that, at fine spatial scales, the relationship between exposure to crime and disadvantaged group memberships is *subadditive*, implying that the sources of disadvantage that are associated with group memberships overlap. This runs counter to the common intuition that disadvantage is compounded across social categories, but is mechanistically sensible: while many things can lead to e.g. living in poor housing, or having a large number of potential offenders nearby, once one acquires such a source of disadvantage there is a limit to how much additional impact it can have. Thus, the sources eventually saturate, with diminishing marginal effects. This nonlinear effect is lost when data is aggregated to the tract level, as would be necessary without the ability to impute at the block level.

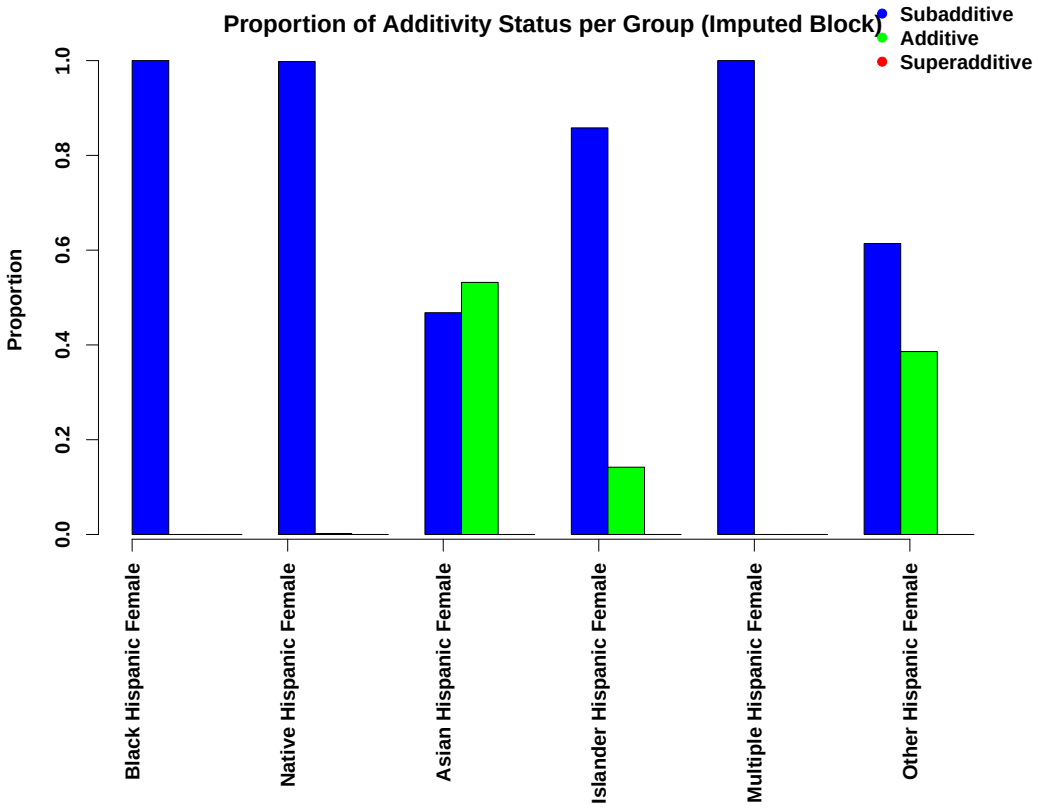


Figure 3.7: Additivity Indices for the three-way block coefficients. Coefficients are in the additive category when the three-way coefficient is less than 5% of the other combined effects.

While the pattern of additivity indices is different between the block and tract level analyses, we note that the general pattern of the coefficients matches. We do not therefore see radical differences across scales, but

rather subtle variations that can be obscured by averaging. As noted, however, those variations can lead to distinct substantive conclusions about that nature of disadvantage in crime exposure.

3.6 Discussion

We have implemented and tested an imputation framework for nested areal units, showing that it produces high quality data for three-way arrays that contain count data. The algorithm specified in this paper should produce high quality data for arrays in which all entries are non-negative integers. Likewise, we are able to leverage data at a higher level of geographic aggregation to optimize the configuration of an imputed array to what we expect the correlations between array cells to be.

With the recent push in many fields within the social sciences for measures constructed at smaller geographic scales, along with the limited availability of some data at such small geographic scales, our imputation algorithm may be applicable in a range of settings. For example, while our case study showed a generally similar pattern in crime exposure for residents of different demographic groups whether measured in census tracts or the smaller spatial unit of blocks, we nonetheless saw sharper and stronger patterns when using the smaller geographic units. Given the spatial segregation of residents across the landscape at varying spatial scales, measuring such effects at smaller geographic scales is arguably substantively important for addressing such research questions. The spatial averaging that occurs when aggregating to larger geographic units has the risk of obscuring such patterns that we were able to observe after imputing the data to blocks.

While the implementation of this algorithm represents a substantial step forward in three-way array imputation, especially with the constraints we describe above, the problem of imputing high order arrays is still difficult. For example, while Bunea and Besag [2000] claim that for the three-way crosstab imputation problems without two-way marginal constraints (i.e. with independent one-way margins), the Monte Carlo method is amenable, this remains a scenario where the formulae are not yet derived and the algorithms are not yet implemented. They also point out that their algorithm is specific to the three-way array case, and while the basic move could, in theory, be adapted to a higher order array, the problem cannot be solved in a "plug and play" fashion. A new transition set must be derived for higher-order arrays, as the transition rule used here is specific to the three-way case (and, in general, each order requires a new set of basic moves, associated with its respective symmetry group). Integrating higher order marginal constraints into models for simulating high-order array data would also be a valuable next step in this line of research.

Another open question in this area is the provable irreducibility of the underlying Markov chain used for the multiple imputation case; while the construction algorithm provided here is verifiable, and the optimization and imputation algorithms guarantee that the result is margin-preserving, the basic move of Bunea and Besag has not been proved to be irreducible in all cases. Their original paper proves irreducibility for any $I \times J \times K$ array such that one of the dimensions has cardinality 2. Irreducibility is thus ensured for the cases studied here, or any other population data using e.g. a two-class sex tabulation. Subsequent work by Lee [2018] provides a proof for the $3 \times 3 \times K$ case, as well as simulation studies suggesting that the property is preserved for cardinalities $4 \times 4 \times 4$ and higher. It thus appears likely that the property holds for all three-way arrays, though this is still unproven. Failure of irreducibility would imply that the Markov chain would not explore the entire state space of possible arrays, thus possibly (1) finding a point estimate that suboptimally captures three-way correlations, or (2) in the multiple-imputation case, providing an imperfect approximation to the target array distribution. (It would not, however, lead to invalid imputations.) Care should thus be exercised when using this method for multiple imputation on arrays that violate the cardinality conditions, when a high level of precision is required.

We observe that the model that we use in this paper to simulate three-way array data is highly scalable. For small areal unit estimation, our imputation scheme does not require information on adjacent units imputed values, so each areal unit can be estimated independently. This has substantial computational benefits, as large sets of areal units can be imputed in parallel, which provides a significant decrease to overall runtime. We were able to simulate three-way distributions for all tracts in the United States, as well as all census blocks in California in less than 24 hours, showing that the algorithm can be employed in over very large regions. As both the construction and optimization problems contribute to runtime, it is likely that arrays with lower population would increase imputation speed, as the construction algorithm will converge more quickly with fewer people.

We note that release of small areal unit data often reflects a "tug-of-war" between advocates of openness, transparency, and data quality (on the one hand) and privacy (on the other). Each faction cites a range of arguments in its favor (often with a certain degree of zealotry), and we here limit ourselves to commenting on implications for the imputation problem. Three-way imputation applied to valid data may or may not allow data identification at a given level of confidence, depending on cell counts; for the scenario studied here, high-confidence identification must come from the two-way marginals, as the higher-order correlation structure is both estimated and approximate. Techniques such as differential privacy can be employed by data collectors to design perturbed marginals that provide guaranteed bounds on identifiability, and algorithms

such as those shown here may be useful for verifying the results of such constructions (and ensuring that they still lead to valid arrays). Similar validation applications are possible for privacy-preserving techniques based on e.g. areal unit aggregation (where units are merged until they no longer permit identification beyond the specified level of confidence). Given that data-perturbing methods like differential privacy pose significant data quality concerns, another use for imputation methods of the type shown here is to ensure that the perturbed data yields imputed arrays that are still appropriate for downstream analysis. Particularly given the importance of neighborhoods, blocks, and other small units for social processes related to social disadvantage, we observe that obfuscation methods that induce systematic bias in small scale structure have the potential to negatively affect policy relevant research impacting vulnerable communities. It is hoped that obfuscators will leverage imputation and related methods to help verify that their modifications will not have such downstream effects.

The imputation techniques introduced here could also be extended in several ways. First, we consider the case where there are additional spatial dependencies amongst the areal units. For this case, the algorithm could be extended by generating a target distribution from both the areal unit immediately higher in the spatial hierarchy from the target unit, but also from that unit's neighbors. A natural approach is to estimate τ_{ijk}^{H} using a spatial smoother at the level of higher-order units, then employing this (rather than the τ_{ijk}^{H} based only on the parent unit) for lower-level imputation. Directly incorporating autocorrelation at the lower level is also possible, but would require a more complex, multi-level and would be less amenable to parallelization. Both are potentially fruitful directions for further work.

Likewise, this technique could also be extended to the case in which the areal units are not perfectly hierarchical. The most obvious direction to extend the algorithm would be similar to the extension described for using multiple parent geographies. In this case, it may make sense to average the correlation structure of parent geographies that overlap with the target areal units. A more radical proposal of this type is suggested for cases in which complete three-way information is available for some units, while only two-way information is available for others. In this case, an interesting option is to train a kernel learner [Scholkopf and Smola, 2001] or similar predictive algorithm to predict the lower-level τ_{ijk} coefficients from observed marginals and other spatial and contextual data; the trained algorithm can then be employed to predict $\hat{\tau}_{ijk}$ directly, as opposed to using τ_{ijk}^{H} as a proxy. Although kernel learning suggests itself due to its interpretation in terms of a similarity function, other methods could be used as well.

On a final, substantive note, we observe that our sample application to exposure-to-crime data suggests that disadvantage in this context is largely subadditive: notably, we do not see the superadditive effects often

presumed (but less often tested) in sociological discussions of intersectionality. We also observe that this subadditivity is largely masked at higher levels of geography (though we do not see evidence of superadditive effects there, either). While it is possible that this is peculiar to the case of crime exposure, the mechanistic interpretation discussed here would suggest that the phenomenon may be much more common. A more systematic investigation of when and how often disadvantage is additive, subadditive, and superadditive across different contexts and for different types of disadvantage would greatly illuminate theory in this area, and may also inform policy interventions. Regardless, our findings reinforce the value of fine-grained spatial data for accurate assessment of local social processes.

3.7 Conclusion

We here specified and demonstrated an algorithm for imputing three-way array data within a hierarchically nested context. This imputation problem is challenging, as it is constrained by the two-way marginal structure of the array, an integer constraint, as well as needing to be optimized with respect to higher order array data. We provide a scalable, robust technique to impute these three-way arrays that relies on Markov Chain Monte Carlo and simulated annealing strategies.

In a test imputation of all tracts in the United States, simulated data from our algorithm produced remarkably low error rates. At the tract level, we observed a mean allocation error of approximately 0.8%, with nearly all tracts having errors below 2.5%. Such errors are better than or comparable to error levels in the Census itself [Khubba et al., 2022], suggesting that imputation is unlikely to be a dominant source of error in subsequent analyses. Likewise, in a case study that examines three-way categories exposure to crime, we found that both imputed data and observed data produced similar conclusions about the relationship between disadvantaged category membership and exposure to crime. Combined, both of these metrics show that the imputed arrays are very similar to observed arrays, and can be used in downstream analyses without introducing significant error.

As sketched above, there is considerable room for further work on the imputation of higher-order array data embedded in spatial hierarchies. With the proliferation of these nested data structures, methods that allow for the use of data at low levels of geographic aggregation where data may be incomplete are particularly valuable.

Chapter 4

Models of Networks with Joint Vertex and Edge Dynamics Containing Arbitrary Dependence

4.1 Introduction

Modeling the time evolution of social network structure remains an active area of theoretical and methodological research. Perhaps the most basic paradigm for such analyses is the network time series, in which one observes a series of cross-sectional snapshots representing the structure of an evolving network at discrete points in time; such a series may represent a discretization of an underlying continuous time process, or in other cases may represent networks whose existence is only constituted at discrete time points (e.g., daily interactions among traders in a marketplace [Baker, 1984]). Considerable progress has been made on models for evolving edge structures in such discrete contexts [see e.g. Hanneke and Xing, 2007, Desmarais and Cranmer, 2012, Krivitsky and Goodreau, 2019, Krivitsky and Handcock, 2014, Mallik and Almquist, 2019]. Many of these models are able to parameterize networks with simultaneous dependence, i.e., the phenomenon in which the presence of one edge in the network can be dependent on the presence (or absence) of another edge, *at the same time point*. Edgewise dependence can be both substantively important and useful for accurate prediction of properties such as triadic closure, and in a cross-sectional setting has been

a factor in the success of the exponential family random graph modeling (ERGM) framework [Lusher et al., 2012]. While often flexible in expressing heterogeneity and/or dependence among edges, however, the above models are restricted to edge dynamics on a fixed vertex set. In other words, the network boundary (and subsequent node set) for a dynamic network must be constant, and the nodes that are present at any point in the network must always be present in the network.

While many social networks of interest can be well-captured by the fixed vertex set assumption, other networks are in *demographic exchange* with their environment. Demographic exchange occurs when individual elements enter and/or leave the network over time (possibly in a fashion that is endogenous to edge structure). For example, the interaction networks formed by attendees at regular meetings of a club or other voluntary organization will show variation as individuals join or leave (or simply are absent on particular occasions). Clearly, such vertex dynamics must affect edge dynamics, if only because edges are by definition possible only for vertices that are co-present; by turns, however, edge dynamics may also affect vertex dynamics. For instance, individuals finding themselves marginalized within interaction networks may be more likely to exit, while those embedded in tight-knit groups may be more likely to persist. Decisions to exit may themselves trigger exits of those similarly situated in the network [Krackhardt and Porter, 1986], while entry of one individual or group may trigger associates to follow (as seen e.g. in the context of disaster response [Butts et al., 2012c]). Capturing such phenomena requires a modeling framework that can accommodate joint evolution of edge and vertex sets.

Perhaps due to the complexity of this joint modeling problem, there has to date been limited work on models for networks with demographic exchange. The first general framework for models of network time series with endogenous vertex dynamics was introduced by Almquist and Butts [2014], who combined a dynamic logistic regression model for vertex presence/absence with dynamic network regression (DNR) temporal exponential-family random graph model (TERGM) to model edge/vertex co-evolution. In this framework, the vertex set at any given time point is approximated as being drawn from a known risk set of potentially active nodes, with the presence or absence of each node being a Bernoulli trial whose success probability may be a complex function of covariates and/or past network states. The edge structure for a given time point is then an inhomogeneous Bernoulli graph conditional on the nodes that are present (as well as the past history of the network, and any covariates). Inference for the resulting model then reduces to a separable pair of logistic regression models, making it tractable even for large systems. The price for this simplicity is the need to assume independence of both edges and vertex co-presence within time steps (though complex dependence on the *past* is possible). This work was subsequently extended by Arabshahi et al. [2015], who employed a

nonparametric latent tree structure for the vertex dynamics (thus allowing some degree of dependence in vertex presence/absence) while retaining the other modeling assumptions. To date, models with parametric vertex dependence specifications, or that combine vertex dynamics with edgewise dependence, have not been investigated.

In this paper, we address this gap by introducing a parametric family of models for joint edge/vertex evolution with dependence in both the edge and vertex sets. Our approach, which builds on past exponential family work (TERGMs and generalized location systems (GLSs)), retains the convenient separable structure introduced by Almquist and Butts [2014], while relaxing the independence requirements. As we show, inference for this family reduces separably to respective pooled ERGM and pooled GLS problems, for which existing inferential strategies can be employed; simulation can likewise by performed by sequential use of existing Markov chain Monte Carlo (MCMC) strategies. While existing TERGM specifications can often be used for the edge process, specifying vertex set dependence requires a new family of statistics. We introduce a number of terms for this purpose, and discuss how they can be employed to capture social dynamics. We also discuss model assessment strategies, demonstrating our approach by application to the classic Freeman et al. [1988] Windsurfer data set.

4.2 Background

As noted, our approach builds on several existing frameworks for modeling systems with complex dependence. We briefly discuss these here, before turning to a more detailed description of our own modeling scheme.

4.2.1 ERGMs and TERGMs

Exponential family random graph models (ERGM) are a parametric family of statistical models that represent the global structure of a network as arising from a set of social forces, operationalized in terms of a set of statistics describing structural properties; the associated parameters can be interpreted as the extent to which draws from the model will on average be biased towards higher or lower values of the statistics in question, versus what would be expected under a reference distribution. Explicit use of exponential family forms for modeling networks was first proposed by Holland and Leinhardt [Holland and Leinhardt, 1981], with subsequent elaboration by many authors (see e.g. Lusher et al. [2012] and Schweinberger et al. [2020] for reviews). Extensive work has been done on the specification of ERGM statistics to capture particular

types of structural properties or processes [Pattison and Robins, 2002, Snijders et al., 2006, Yaveroğlu et al., 2015, Lee and Butts, 2020], and the framework has been applied in a wide range of settings both within and beyond the social sciences [e.g. Saul and Filkov, 2007, Schaefer et al., 2017, Goodreau et al., 2009, Daraganova et al., 2012, Krivitsky and Morris, 2017, Yu et al., 2020]. Formally, we say that a random graph G on support $\mathcal{G}$ is expressed in ERGM form when its probability mass function is written as

$$\Pr(G = g|s, X, \theta) = \frac{h(g)\exp(\theta^T s(g, X))}{\sum_{g' \in \mathcal{G}} h(g')\exp(\theta^T s(g', X))}, \tag{4.1}$$

where $s : \mathcal{G}, X \mapsto \mathbb{R}^p$ is a vector of sufficient statistics, $\theta \in \mathbb{R}^p$ is a parameter vector, X is a covariate set, and $h : \mathcal{G} \mapsto \mathbb{R}_{\geq 0}$ is a reference measure (which determines the behavior of the model as $\theta \to 0$). In typical applications, $\mathcal{G}$ is taken to be the set of all graphs or digraphs on a fixed vertex set, V, yielding a cross-sectional model for the edge structure of a network, given its composition. Likewise, G is usually taken to be unvalued, although valued generalizations are also possible [see e.g. Cranmer and Desmarais, 2011, Krivitsky, 2012, Krivitsky and Butts, 2017].

Given a network time series, it is natural to specify a model family in which each network in the series is drawn from an ERGM, conditional on the past networks (i.e., taking past states as a covariate). Such models are called temporal ERGMs, or TERGMs. Formally, given random graph series $\ldots, G^{(t-1)}, G^{(t)}$, the TERGM pmf at time t is given by

$$\Pr(G^{(i)} = g|s, G^{(<t)}, X, \theta) = \mathrm{ERGM}(g|s, X \cup G^{(<t)}, \theta); \tag{4.2}$$

as this specification is temporally causal, the joint pmf of multiple time points is then the product of the above taken at each. Typically, TERGM specifications are distinguished by statistics that are (possibly complex) functions of past states, allowing edge states in the present time step to be influenced by past network states; importantly, since the past is a covariate to the present, such influence can be modeled without assuming dependence in the current time step. TERGMs without such dependence are said to be "dynamic network regression" or DNR TERGMs, as they reduce in form to logistic regression. For networks with short intervals between measurements, DNR TERGMs can potentially perform well, since dependence on the past accounts for much of the entanglement in current edge states [Almquist and Butts, 2013]; this performance generally degrades with larger time intervals, however, as within-time step dependence becomes a more substantial contributor to network evolution [Lerner et al., 2013].

The TERGMs are an extremely flexible framework for network time series modeling, with the ability to

specify temporal dependence, dependence in the present, and complex covariate effects in a highly general way. The obvious limitation of TERGMs, from our standpoint, is the assumption of a fixed vertex set. Relaxing this requires use of a different distributional formalism, to which we now turn.

4.2.2 Generalized Location Systems

While there has been a long history of the development of random graph models, there has been comparatively less development for models examining vertex dynamics. One framework of this type is the Generalized Location System (GLS), which is a family of models for the assignment of arbitrary entities (people, objects, organizations) to generalized "locations" (which may be physical locations, states, etc.) [Butts, 2007]. Like the ERGMs, the GLS has an exponential family form, with the random variable being a state vector, $S \in \mathcal{S}$, such that S_i is the index of the location to which the ith object is assigned. The GLS pmf is then

$$\Pr(S = \ell | r, X, \lambda) = \frac{u(\ell) \exp(\lambda^T r(\ell, X))}{\sum_{\ell' \in \mathcal{S}} u(\ell') \exp(\lambda^T r(\ell', X))}, \tag{4.3}$$

where $r : \mathcal{S}, X \mapsto \mathbb{R}^k$ is a vector of sufficient statistics, $\lambda \in \mathbb{R}^k$ is a parameter vector, X is a covariate set, and $u : \mathcal{S} \mapsto \mathbb{R}_{\geq 0}$ is a reference measure. By appropriate choice of r, the GLS can capture not only covariate effects (e.g., tendencies for specific individuals to be found in specific locations) but also dependence among assignments (e.g., tendencies for certain individuals to be found in the same locations, for certain locations to be occupied by individuals with similar characteristics). Although the GLS as defined is cross-sectional, we may also extend it to the dynamic case in the same manner as the TERGMs. I.e., given a sequence of state vectors $\ldots, S^{(t-1)}, S^(t)$, we may define the conditional distribution

$$\Pr(S^{(t)} = \ell | r, S^{(<t)}, X, \lambda) = \text{GLS}(\ell | r, X \cup S^{(<t)}, \lambda), \tag{4.4}$$

leading to a "temporal GLS" or TGLS family analogous to the TERGMs, and applicable to discrete time dynamics.

While the GLS was originally developed for modeling e.g. occupational stratification and residential settlement patterns, it can also be used to model vertex dynamics. We accomplish this by defining two "locations," one that represents presence within the network, and one that represents absence from the network. This allows for a model of assignments to locations to represent participation in the network at a given time point. Appropriate choice of statistics then allows for dependent participation (i.e., the presence of one

vertex may depend on the presence or absence of others), while past states of the networks may also enter in as covariates.

4.2.3 Endogenous Vertex TERGMs

As noted above, our development builds on the endogenous vertex TERGMs proposed by Almquist and Butts [2014]; in terms of the above, these models are mathematically equivalent to DNR TERGMs combined with an independence TGLS, with the hierarchically specified structure

$$\Pr(G^{(t)}=g,S^{(t)}=\ell|s,u,G^{(<t)},S^{(<t)},X,\theta,\lambda)=\mathrm{ERGM}(g|s,X\cup G^{(<t)}\cup S^{(\leq t)})\mathrm{GLS}(\ell|r,X\cup G^{(<t)}\cup S^{(<t)}). \tag{4.5}$$

Conceptually, we may think of this family as describing a process where, at each time point, the vertex set $V^{(t)}$ is drawn from an independence GLS conditional on past network states, and then the edge set $E^{(t)}$ is drawn from an independence ERGM conditional on both $V^{(t)}$ and the past history of the network. The critical simplifying condition here is the assumption that neither s nor t contain dependence terms, which amounts to the approximation that any dependence in the current time step can be accounted for by some combination of covariate effects and past history. As noted, we here relax that assumption, allowing for a more general treatment.

The family of Arabshahi et al. [2015] is also closely related to Eq. 4.5, although the vertex process involves a latent variable model that is not precisely equivalent to a GLS; it can however be thought of as an "extended" GLS featuring "virtual nodes" that do not actually appear in the network, but that influence the appearance or other (manifest) nodes. Dependence is allowed, but the dependence graph on vertex states is restricted to acyclic pairwise (i.e., tree) structure. Although limiting, this scheme is computationally efficient, and allows for certain important cases (e.g., latent groups that tend to all appear together). We do not pursue this further, here, focusing entirely on the parametric case.

4.3 Joint Modeling Framework

Here, we summarize the joint vertex/edge modeling framework used in the remainder of the paper. Parameterization issues are then discussed in Section 4.3.2, inference in Section 4.3.3, and simulation in Section 4.3.4.

4.3.1 Model Formalism

As described in Section 4.2.3, our approach generalizes that of Almquist and Butts [2014] by relaxing independence assumptions. For clarity in the present application, we modify our notation slightly from that of Eq. 4.1 and 4.3, though the basic framework is (net of relaxations) equivalent to that of Eq. 4.5 above.

In what follows, we will take $G^{(t)} = (V^{(t)}, E^{(t)})$ to be a (random) graph observed at time t, with V and E representing the (also random) vertex sets. We assume that $V^{(t)} \in \mathcal{V}$ for all t, for some fixed set of possible vertex sets $\mathcal{V}$ (i.e., the support of the vertex process is fixed and known). We also define $\mathcal{G}_V$ to be the set of all graphs or digraphs (as desired) on vertex set V.[1] Let $G_{(i)}^{(i+k)} = (G^{(i)}, G^{(i+1)}, \ldots, G^{(i+k)})$. Then we define the lag-k pmf of the joint edge/vertex process with realization $g = (v, e)$ to be given by

$$\Pr(G^{(t)} = g | s, r, G_{(t-k)}^{(t-1)}, X, \theta, \lambda) = \frac{h(g)\exp(\theta^T s(g, X \cup G_{(t-k)}^{(t-1)}))}{\sum_{g' \in \mathcal{G}_{V^{(t)}}} h(g')\exp(\theta^T s(g', X \cup G_{(t-k)}^{(t-1)})} \frac{u(v)\exp(\lambda^T r(v, X \cup G_{(t-k)}^{(t-1)}))}{\sum_{v' \in \mathcal{V}} u(v')\exp(\lambda^T r(v', X \cup G_{(t-k)}^{(t-1)}))}, \tag{4.6}$$

with statistics r and s, reference measures h and u, parameters θ, λ, and X defined as above. We do not assume any particular dependence form for r or s.

Now, consider an observed sequence $g^{(0)}, \ldots, g^{(t)}$. We then work with the conditional likelihood

$$\Pr(G_{(k)}^{(t)} = g_{(k)}^{(t)} | s, r, g_{(k-1)}^{(0)}, X, \theta, \lambda) = \prod_{i=k}^{t} \Pr(G^{(i)} = g^{(i)} | s, r, g_{(i-k)}^{(i-1)}, X, \theta, \lambda) \tag{4.7}$$

$$= \left[\prod_{i=k}^{t} \frac{h(g^{(i)})\exp(\theta^T s(g^{(i)}, X \cup g_{(i-k)}^{(i-1)}))}{\sum_{g' \in \mathcal{G}_{v^{(i)}}} h(g')\exp(\theta^T s(g', X \cup g_{(i-k)}^{(i-1)}))}\right] \left[\prod_{i=k}^{t} \frac{u(v^{(i)})\exp(\lambda^T r(v^{(i)}, X \cup g_{(i-k)}^{(i-1)}))}{\sum_{v' \in \mathcal{V}} u(v')\exp(\lambda^T r(v', X \cup g_{(i-k)}^{(i-1)}))}\right] \tag{4.8}$$

$$= \left[\prod_{i=k}^{t} \mathrm{ERGM}_{v^{(i)}}(g^{(i)} | s, X \cup g_{(i-k)}^{(i-1)}, \theta)\right] \left[\prod_{i=k}^{t} \mathrm{GLS}(v^{(i)} | r, X \cup g_{(i-k)}^{(i-1)}, \lambda)\right], \tag{4.9}$$

where we use the notation ERGM_V to specifically denote an ERGM with support on $\mathcal{G}_V$. We observe that the joint likelihood factors in Eq. 4.9 into two terms, one of which is a product of ERGM pmfs (and depends only on θ), and the other of which is a product of GLS pmfs (and depends only on λ). It follows that the conditional likelihood of the edge parameters, θ, has the form of a pooled ERGM, while the conditional likelihood of the vertex parameters, λ, has the form of a pooled GLS, and the two are *inferentially separable*. We exploit this property in Section 4.3.3.

[1] Additional support constraints can also be incorporated, but are not treated explicitly here for economy of notation.

While the vertex and edge processes are inferentially separable, it should be emphasized that they are not *predictively* separable; that is, future realizations of the vertex process directly affect subsequent draws from the edge process, and vice versa. The two processes iteratively feed back on each other, making it possible to capture phenomena in which relational changes lead to demographic changes, which in turn drive further relational dynamics. We discuss simulation from the joint vertex/edge process in Section 4.3.4.

4.3.2 Parameterization

As noted above, parameterization of the joint edge/vertex model is controlled by the respective choices of sufficient statistics s and r, and secondarily by selection of the reference measures h and u. As the two processes are parameterized separately, we discuss each in turn.

Edge Process Parameterization

As the edge process for the joint model is of the same form as a standard TERGM, most of the usual tools and principles for TERGM parameterization carry over directly [Block et al., 2022]. Additionally, the inclusion of a separate vertex model would allow for some new terms that depend on the vertex processes. For example, one could parameterize a new tergm term that describes the past participation of a node (at a given time lag), and use this past behavior as a predictor of tie probabilities in the present.

The primary distinct consideration for edge process parameterization with vertex dynamics is the choice of reference measure. The counting measure ($h(g) \propto 1$), often used as a default choice in ERGM inference, implies that baseline mean degree will scale proportionately with graph size (i.e., constant baseline density). Krivitsky et al. [2011] introduced an alternative measure ($h(g) = 1/|V|^{t_e(g)}$, where t_e is the edge count) that leads to asymptotically constant mean degree, which is often more realistic for social networks of moderate to large size. This was further generalized by Butts and Almquist [2015], who proposed the reference $h(g) = |V|^{\gamma-1}$, where γ is a user-selected parameter. This takes the counting measure ($\gamma = 1$) and the Krivitsky reference ($\gamma = 0$) as special cases, with other values of γ yielding baseline mean degrees that scale as power laws in $|V|$; this can be used to capture e.g. the somewhat misleadingly named "power law densification" observed by Leskovec et al. [2007], in which mean degrees in some networks scale sublinearly with network size. (Such networks do in fact become sparser as they grow, but they do so more slowly than a classical "sparse graph" with constant expected degree.) Although motivated on other grounds, Almquist

and Butts [2014] also employ a parameterization equivalent to the flexible reference measure, in which γ is endogenous: they simply include the statistic $t_e(g) \log |V|$ in the model as a standard term, otherwise using the counting measure. This allows mean degree scaling to be estimated directly from the data, rather than being specified *ex ante*.

Vertex Process Parameterization

In parameterizing the vertex component of this model (a selection of effects for r), we leverage work completed by prior literature [Almquist and Butts, 2014]. This work describes a set of terms that match a set of social processes. While prior literature couches the development of model terms in the context of modeling the Freeman et. al windsurfer data [Freeman et al., 1988], the processes that are described by these terms are still useful in a broader social context. We present a quorum of terms here that extend the terms presented in previous literature. We structure this section by examining the social processes that would be described by each model term.

Vertex Model Notation For this section, we use the following notation. Our goal is to specify the set of statistics $r(v', X \cup G^{(t-1)}_{(t-k)})$. With two states, state 1 is being present in the network, while state 2 is being outside of the network at that time. a given statistic would be defined as r_i. We define the covariate values for nodes as X and the covariate values for graphs as X^G. Graph covariate values are applied consistently across the risk set, with R entries per graph level covariate value (i.e. the value of the covariate is duplicated for all nodes). For the size of the risk set, we use R. We model $M - K$ time points. Q_V is a vector the length of the number of nodes in the risk set. This vector has a 1 if a node is participating, and zero otherwise. Statistics are defined per time slice of the network. For the first set of terms here, we also define a vector B that includes information on the vertices present, with it's exact specification defined in each section.

Average Network Size Much as we use an edges term in the ERGM context, we may be interested in describing the average size of a network on a given day. For this, we would use a network size term. This term would describe the baseline likelihood that any node participates in the network. The statistic associated with this term would be be the number of people present in the network on a given day. Mathematically, it is equivalent to $r_s = \Sigma_{i=1}^{R} B_i Q_{V_i}$. B is simply a vector of 1s for this term. This term is slightly different than the other terms that we present here, as it can act as a model intercept. Models that only contain the network size term will be guaranteed to reproduce the average network size (across all networks).

Inertia Effects The inertia effect we describe here is similar to the one described by [Almquist and Butts, 2014]. This term adds a statistic equal to the difference of the number of nodes in the network at time $t-k$ and the number of nodes not participating in the network at time $t-k$. Equivalently the statistic for this term is: $r_i = \Sigma_{i=1}^{R} B_i Q_{V_i}$. B is a vector containing a 1 if a node participated in the network at the previous time point, and a -1 otherwise. This term describes a persistence effect, where if the coefficient is positive, there is a higher likelihood that a node will be in the same state at the next time point. The persistence effect described here is symmetric, where a positive coefficient would inhibit new entries, as well as encouraging continued vertex participation.

Exogenous Node Level Covariate Effects Node level covariates can be integrated into this model using a statistic for each covariate (if the covariate is binary or continuous). This effect would add a statistic equivalent to the sum of the covariate values for nodes present in the network. Mathematically, we can represent this statistic as $r_c = \Sigma_1^R X_i Q_{V_i}$. For binary covariates, such as group memberships, this term would describe the effects of group membership on the likelihood of network participation (and could incorporate a lag of zero). In the case of continuous covariates, this statistic assumes that the difference between values of a covariate are constant. For a categorical variable, several binary dummy variables would need to be generated for model inclusion using this term.

Some of the regularity and individual effects that are described in prior literature [Almquist and Butts, 2014] would be represented as this effect type in this model framework. Likewise, if one wants to include the past as a covariate to the present (i.e. include triadic embeddedness or degree effects from previous networks as a predictor for network participation), covariates could be specified for nodes that contain these data. In this case, X would contain information on the past condition of nodes in the risk set.

Exogenous Graph Level Covariate Effects Given that we model a series of graphs, there may be cases in which graphs have covariates of interest attached to them (e.g. days of the week, environmental covariates). To include these in the model, we introduce a statistic for each binary or continuous variable of interest. This statistic will be equal to the sum of the covariate value for each graph multiplied by the number of nodes present. Mathematically, this statistic is written as $r_c = \Sigma_1^R X_i^G Q_{V_i}$. In other words, the graph level covariate will evenly affect participation likelihood for all potential nodes.

A use for this type of effect is the seasonality effects described by prior literature. Parameterizing a model that includes effects for days of the week, seasons, or months would include this type of seasonality in a

model.

Dyadic Copresence Copresence in the present can be a powerful predictor of network participation. In other words, the likelihood of a node's participation could be based on the participation of other nodes at the same time point, assuming that nodes coparticipated in the network in the past. We introduce a single statistic that describes this effect, which is equivalent to $r_d = \Sigma_{i=1}^{R}\Sigma_{j=1}^{R} A_{t-k_{ij}}|Q'_{V_i} - Q'_{V_j}|$. Here, A is a matrix describing the coparticipation behavior of nodes at time $t-k$. This matrix has a 1 at the i, j cell if i and j both are in the network at time $t-k$. Q' is a binary vector, where an entry in this vector is 1 if that node is out of the network, and zero otherwise. The participation vector describes the behavior of a node at time t. Combined, this term will have a negative coefficient if nodes who co-participated in the past tend to both be in or out of the network in the present. If past patterns of co-participation tend to encourage only one of the nodes to participate, then this term will have a positive coefficient.

Attributional Dyadic Copresence While generally, dyadic copresence may function across all nodes in the network, one may wish to have the dyadic copresence effect separated by some covariate value. For example, if there are multiple subgroups present in the network that have active boundary maintenance occurring, the effect described by dyadic copresence may be subgroup specific. For this case, we introduce a single statistic that is equal to $r_d = \Sigma_{i=1}^{R}\Sigma_{j=1}^{R} A_{t-k_{ij}}|Q'_{V_i} - Q'_{V_j}|$. The difference from the dyadic copresence term described above is in the A matrix, which will only have a 1 for coparticipation if both nodes belong to the same group. This is a special case of the generalized dyadic copresence term that is described above.

4.3.3 Parameter Estimation

Given the model formalism and set of potential effects, we next discuss how to estimate coefficients and fit the model. To fit the model, we rely on the result shown above (see section 4.3.1), which shows that the model can be fit in a separable manner. This also allows us to have different specifications for each process, and more importantly, fit each component independently. This separability assumption allows us to use existing tools and procedures to fit each component of the model. However, there is still some accounting that must be completed to keep track of terms.

In the event that there are no dependence effects in the model, fitting these models is trivial, and can be done using the MPLE. This is due to the independent nature of the model, which allows for the probability of the

presence of a node or edge to be evaluated independently. In modelling terms, these models are equivalent to logistic regression models in this mode. However, when there are dependence effects parameterized in either the vertex or edge models, additional modelling techniques are required to fit the models. The reason that the MPLE cannot be used to accurately model edge probability or vertex participation probability is due to a given edge being dependent on other edges at the same point in time, violating the i.i.d. assumptions. The MPLE will converge to the MLE if and only if the model specified is an independence model (this does not indicate that the MPLE cannot be useful, or approximate the MLE in some cases [Hummel et al., 2012]). In the event that dependence term are present in the model, the MLE will provide better coefficient estimates.

In the event that there are dependence effects present in the model, we rely on MCMC to estimate the normalizing factor and fit these models. Given the complexity of the distributions defined by the normalizing factor, MCMC is able to provide estimates for coefficients that converge to the MLE, which functions regardless of the dependence structure of the model. Prior literature has described the ways in which MCMC can be used to get estimates that converge to the MLE [Geyer and Thompson, 1992]. Specifically, we use the Geyer Thompson algorithm to fit both the vertex and the edge components of the model.

To fit the model and estimate coefficients, we use two R packages. The first of these is the location package, which implements a set of tools to fit and estimate Generalized Location Systems. The second package is the ergm.multinet package, which extends the standard ERGM package provided by Statnet [Handcock et al., 2008b] to multinetwork ERGMs, which are equivalent to fitting a conditional TERGM on the set of networks with distinct node sets. For both the vertex and edge processes, we utilize the Geyer-Thompson technique. Model fitting parameters that depart from the defaults include a thinning parameter of 100 for the vertex process, to ensure independent draws from the Markov chain, and an increased number of maximum iterations for the adaptive fitting used by the edge process.

4.3.4 Simulating from the Model

Once a model has been fit, we can leverage the parametric nature of the model to simulate from it. While the model is inferentially separable, the result that is shown in section 4.3.1 highlights that the model framework is not predictively separable. This requires several choices in the design of a simulator from this model. We leverage the simulation strategy used in prior literature [Almquist and Butts, 2014], which assumes that the vertex participation process occurs prior to the edge process at a given time point. In the language of the model, we will first draw from the vertex process, and then draw an edge set given the draw from the vertex

process. This will be referred to as a 'leapfrogging' algorithm in the future, as it allows us to make sequential draws from each of the models. This algorithm is represented in Figure 4.1.

While this simulation strategy is relatively straightforward, there are several considerations for implementation. First, the *risk set*, or the set of nodes that can potentially be in the network at any point in time, must be known. Given that both the vertex and the edge process may require information about the covariates and interaction/participation patterns for the nodes, a finite set that is known prior to inference or simulation is required. A second consideration is that full prediction, especially predicting more than one time step into the future is a difficult simulation target [Mallik and Almquist, 2019]. In this paper, we specifically focus on the one-step prediction/simulation problem, where we simulate networks given the immediate (lagged) past as a baseline for the simulation, for all simulated draws.

This simulator has been implemented in the R computing language using functions from both the ergm and location packages. It has been calibrated to do single step imputation for an arbitrary number of simulation replicates, primarily for use in model adequacy checking.

4.4 Adequacy Checking

Given the network simulation schema described above, we next turn to a procedure for model adequacy checking. We use one-step prediction to evaluate the ability of a given model to reproduce structural features of the network. Specifically, given the state of the network at time $t - k$, we then predict the state of the network at time t. Given the simulated networks, we then compare the values of a set of structural statistics about these networks to observed statistics on the networks. For the purposes of this paper, we follow a similar procedure to the one specified by Almquist and Butts [Almquist and Butts, 2014] in their work on joint vertex/edge processes.

First, from a fitted model, we know that simulated draws from this model will conditionally match the mean target statistics (i.e. parameterized effects) from the model, by nature of the model being exponential family. However, if the model is a valid model for understanding the dynamics of the observed system, then the simulated draws should also match other structural features of the observed network that were not parameterized. These structural features can involve both the vertex and edge features, although if we simulate and compare the edge based structure of the network, this is also capturing information about the vertex dynamics, as any draw from the edge model is conditional on the vertex selection (predictively).

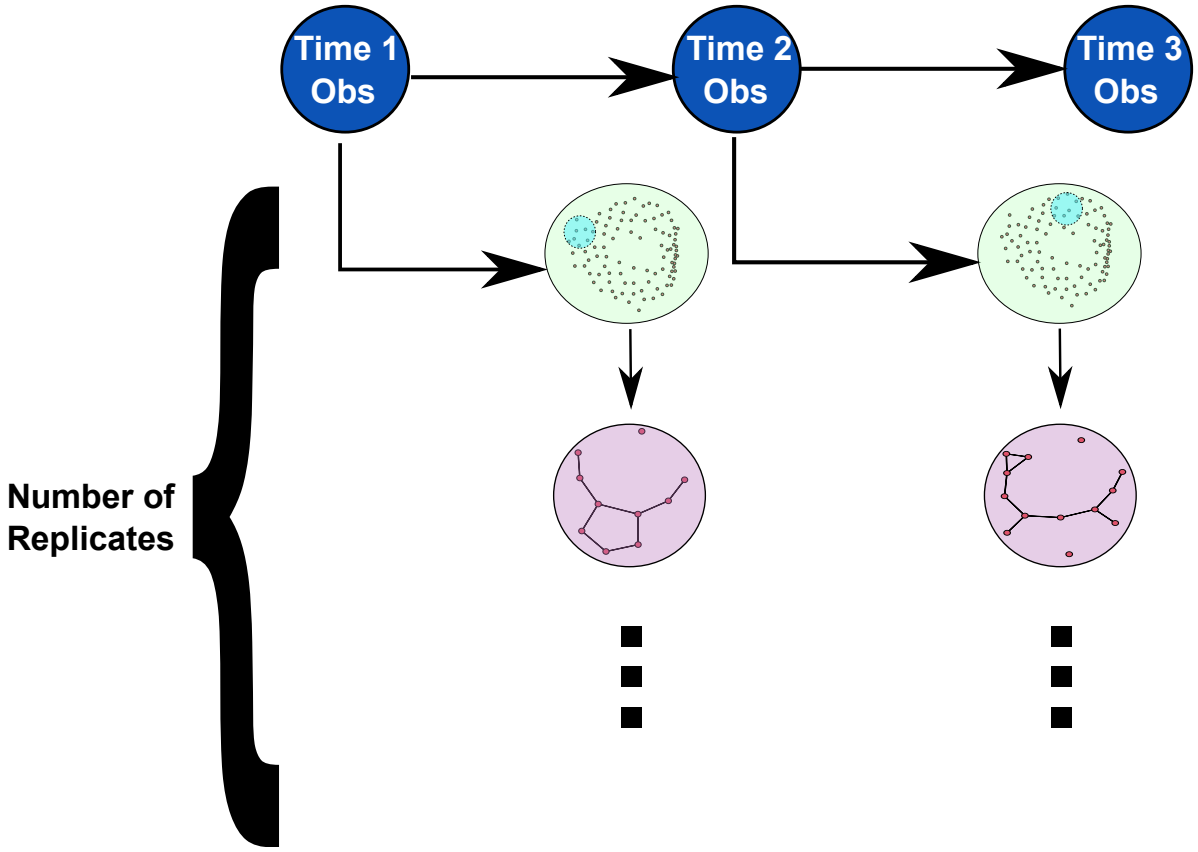


Figure 4.1: A visualization of the strategy used for a one-step simulation process. The simulation begins with a vertex draw based on the state of the network at the most recent time point. We explicitly condition on the first time point, using it to draw a set of vertices for time point 2. From this simulated vertex draw, we then draw a set of edges, which when combined with the vertex draw would be the state of the simulated network at a given time point. If we want multiple network draws, we can simply take m vertex draws using the state of the network at the previous time point and the fitted model, and then take edge draws based on the simulated vertices. Multiple step simulation can also be done. Instead of always starting with the state of the network at the previous time point, we can instead use the simulated network of a previous time point as the starting point for the vertex process at the current time point.

Therefore, to assess model adequacy, we simulate n draws from the model, and compare these draws to the observed network on a set of structural features. We consider a model to be adequate at reproducing structural features of the observed networks if the observed networks data is within the α percentile of the distribution of simulated networks.

For this project, we use one-step prediction in order to assess model adequacy. This algorithm simulates networks for a timepoint t using information based off the observed time point at $t - k$. Then, we compare the distribution of simulated networks at time t to the observed network at time t. This allows for the networks to be comparable, as they are both networks at the same point in time. We also base decisions on

model adequacy on a 95% simulation interval, where we consider a model adequate if the observed networks lie within a 95% simulation interval of the simulated networks.

4.5 Empirical Case

Given our implementation of the joint vertex/edge model, we next fit a series of models and demonstrate model fit. The models we describe in this section are all 1-lagged models, which only include information from the most recent network that was observed.

4.5.1 Data

For this empirical case study, we use Freeman et. al's windsurfer data [Freeman et al., 1988]. These data are a dynamic network containing a month of daily observations of social interactions amongst windsurfers on a beach. The data is encoded as a set of 31 networks, each of which contains the set of social interactions amongst the windsurfers for that day. On most days, a different set of windsurfers are on the beach, meaning that the active nodeset for each network is not constant. In total, there are 95 windsurfers who attend the beach conversations on at least 1 day. There was one day that is missing in the dataset, which was removed for this analysis. Additionally, we condition on the first day of the network, leaving 29 networks to fit to and simulate data for.

In addition to the adjacency matrices that make up the dataset, there is also a set of covariates that are attached to both the nodes and the networks themselves. The covariate set for the nodes includes information about whether each windsurfer is a 'regular' on the beach, meaning that they are frequent attendees and are well integrated into the social communities on the beach. This information was assessed qualitatively on ethnographic grounds [Freeman et al., 1988]. In addition, this category of regulars was separated into two groups of regular attendees, both of which were determined qualitatively. The groups are denoted as 'Group 1' and 'Group 2'. There are 54 windsurfers who are regulars on the beach, while there are 43 people who are classified into group 1 or group 2, with the remainder of the regulars not being classified into either group. Group classifications are exclusive.

Despite there being a group of 'regulars' on the beach, the beach networks are not consistently large compared to the risk set of nodes. On a given day, the number of windsurfers who are present ranges from 3 windsurfers

to 37 windsurfers, with the average network size being 11.5 windsurfers. It also appears that there are seasonal patterns in attendance, with weekends being more popular than weekdays for windsurfing.

The covariates that are attached to the network also include the day of the week that the network was observed, as well as environmental covariates such as the temperature, wind speed, and wave height [Almquist and Butts, 2014]. We will integrate seasonality into the models using the day of week covariates. Environmental covariates can be used to determine the environmental determinants of network participation, which may be integrated into future work.

The locations were constructed as described in section 4.2.2. We consider a windsurfer to have participated in the network on a given day if they are recorded as being on the beach on that day, even if they have no ties to other windsurfers. This classification identifies that isolates would still be in the network and even just copresence on the beach without having present ties can influence the structure of the network. For windsurfers who are out of the network, these vertices were not recorded as being on the beach on the given day. We assume that there are no network ties across the network boundary.

4.5.2 Potential Model Specifications

In this section, we propose a set of model terms that we believe would describe the evolution of the vertex and edge processes that occur on the beach. Many of the effects that we describe in this section have been identified by prior work as potentially important social factors that can determine the evolution of this dynamic social network. In the vertex portion of the model, we identify 9 potential effects. These effects include an effect for the average size of a network, which we hypothesize would constrain the general tendency for windsurfers to be present on a given day.

We include several covariate effects. We hypothesize that group memberships are likely to encourage participation in the network across all points in time. As a result, we include one effect for group membership in the regular category, as well as in each of the ethnographically identified groups. We expect that subgroup membership may also endogenously enhance network participation, which we will discuss in later terms.

From past network participation patterns, we hypothesize that embeddedness in triangles is likely to enhance the likelihood of network participation in the present. We include a term that is equal to the number of triangles a node was present in at the previous time point. The hypothesis behind this term argues that the more embedded in the network structure a windsurfer was at the previous time point, the more likely they

are to be present at the current time point. This hypothesis is based on the idea that those embedded in social structures are likely to maintain membership in those same structures.

We also include an effect for inertia. Given that decision making tends to be affected by inertia [Alós-Ferrer et al., 2016], we hypothesize that if a windsurfer attends the network on a given day, they will be more likely to also attend on the following day. Likewise, if they are absent on a given day, it is more likely that they stay out of the network the following day. In other words, the network boundary is likely to maintain attendance across days, in general.

At the graph level, we include effects for seasonality. We include one effect for each day of the week (omitting Monday as a reference category) as a graph level covariate. Given work commitments and other seasonal effects, we would expect that network participation would be enhanced on weekends and suppressed across the week. This has been previously identified as an important factor in the evolution of this dynamic network.

For dependence effects, we include four terms. We include a general dyadic copresence term, for which we hypothesize that prior coparticipation is likely to enhance present coparticipation. However, given the endogenously maintained group structure present in these networks, we also include terms for dyadic copresence by the regular group, group1 and group2. These terms are likely to enhance group participation for windsurfers who are part of the regular and other subgroups. These terms are also able to leverage the flexible structure of the model, which can parameterize arbitrary dependence.

For the edge model, we use a slightly more simplified set of model terms, drawn from previous literature that models this network [Almquist and Butts, 2014]. These terms include an edges term, nodematch terms on regular, group 1 and group 2 memberships, and a seasonality term (in the form of edgecov matrices for each day of the week except Monday). We also include a term that allows for generalized degree scaling. Given the differences in network size on different days, this term will allow us to have degree scaling by the number of nodes present. Finally, we include a term for triadic closure (gwesp). This term has a fixed decay parameter identified by a preliminary search of the decay values from 0 to 1.

4.5.3 Model Selection

Forward Selection is used to find the best fitting model, using the potential terms described in section 4.5.2. We initialize the forward selection process with the vertex process containing only a size term (i.e. the Bernoulli model), and the edge model with the edges and degree scaling terms. To determine model fit, we

use a structural measure. A structural measure is a reasonable metric to discriminate between models, as it incorporates the effects of the vertex and edge processes on each other. If we were to use a criterion measure (e.g. AIC, BIC, etc.), we examine only the inferential part of the model, for which the vertex and edge processes are separable. A structural measure requires simulation, for which the vertex and edge processes are not separable (see section 4.3.1).

Specifically, we use the hamming distance between the simulated and observed networks. We construct each network using the following schema. First, we generate a network containing all vertices in the risk set, but no edges. We then use the one-step prediction procedure from Section 4.4 to get a simulated network draw. Then, at each time point, we add the simulated network's edges to this empty graph, as well as a loop for each node that is identified as "present" in the network. These self loops are an identifier that allows us to better include the effect of the vertex selection process into this measure, and differentiates between non-participating nodes and isolates. In adding the loops, this is equivalent to adding the Hamming distance for the vertex participation process to the Hamming distance for the edge process. We repeat this process with the observed networks. Finally, we compute the Hamming distance between the simulated and observed networks, which acts as a measure of dissimilarity or an error metric. The final value is the average of the errors at each time point. On testing, we found that this measure did describe the ability of a model to simulate networks like the observed ones, but was very noisy. To minimize the effect of noise in the vertex and edge simulations, we simulate 30 draws for each time point, and take the mean hamming distances of these 30 networks and the observed network. Models in which the fitting procedure fails due to parameterization issues are excluded from this calculation (this only affects the inclusion of the dependence terms in the vertex process, as it is likely that these produce instability).For the cases where there is only one node selected and the case where no nodes are participating in the network, we have several special cases. The isolate case would have a simulated network constructed in which there is only one node with a self-tie, and no other edges present. For the case where there are no nodes selected for participation, then the simulated network is just the initial empty graph described above.

In forward selection, a model term would need to improve the error rate described here in order to be selected. Model selection continues until no terms produce an improvement to the error rate, or all terms are present in the model. If a term is not selected, we interpret this as being indicative of this term not being necessary to produce structurally similar networks to the observed data.

While the number of potential models precludes an exhaustive search of the model space to find the best fitting model (especially if we cannot discriminate purely off the inferential process), we can still utilize

methods like forward selection to search the model space in a systematic way, without having to fit every possible model. We complete this model search using the terms provided in section 4.5.2.

4.5.4 Model Results and Adequacy

The best fitting model identified by the forward selection design is summarized in table 4.1.

Model Results	
Vertex Terms	Coefficient
Graph Size	-2.054***
Tuesday	-0.200
Wednesday	0.125
Thursday	0.319
Friday	0.170
Saturday	1.111***
Sunday	1.106***
Edge Terms	
Edges	-1.371***
Nodematch Regular	0.127*
Degree Scaling Offset	-0.778***
Gwesp	1.083***
Sunday	0.020
Joint AIC	3079.484
Hamming Error	47.083

Table 4.1: Model Coefficients for the best fitting model

This model shows several interesting patterns in the vertex and edge parameters. This section has grouped the terms by context to discuss them.

Group Membership Group membership in the regular subgroup increases the likelihood of nodes sharing an edge. Given the nature of the this group being an enfranchised group of windsurfers within this community,

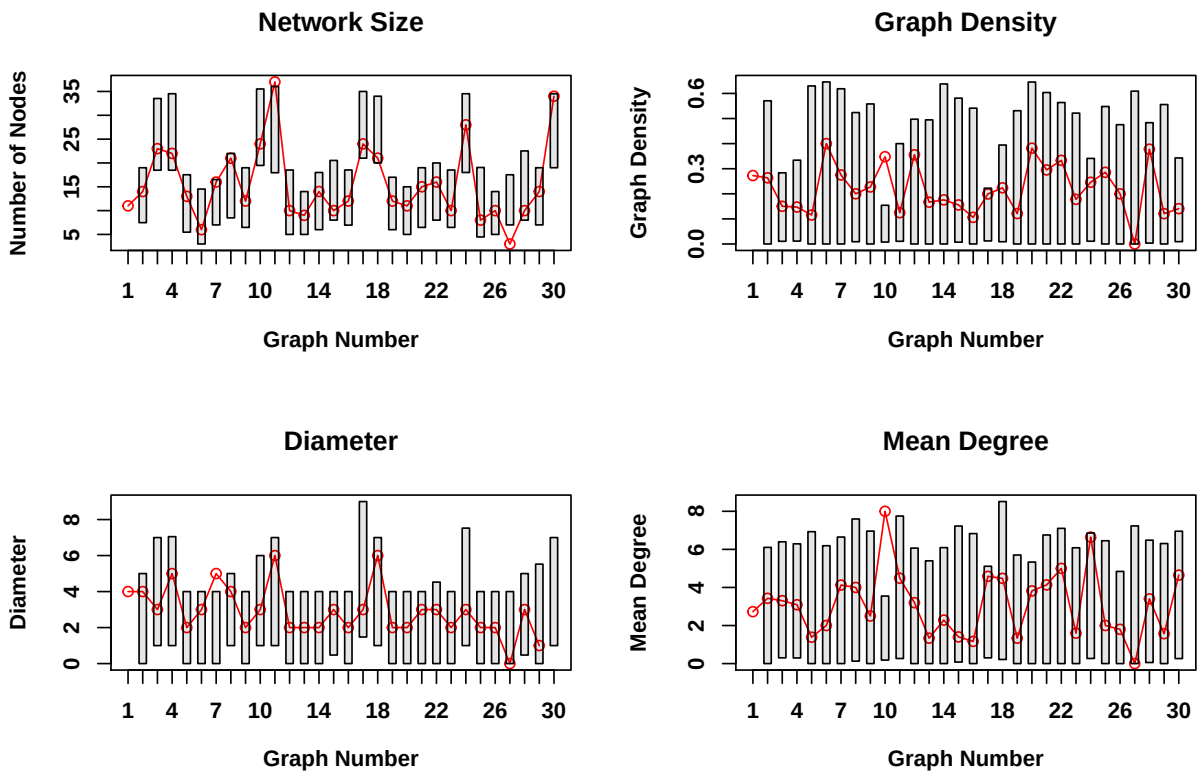


Figure 4.2: Simulated Network Statistics vs. Observed Network statistics. The red line represents the observed network statistic, while the boxplots represent 95% simulation intervals. Any missing time points have been removed. Simulation intervals are based on simulations of 100 networks per time point.

it follows that embeddedness in this community would increase the likelihood of edge formation.

Seasonality The vertex model indicates that most weekdays have a higher likelihood for nodes to be present, while weekends produce a significantly higher chance of windsurfer participation in the network. Given the standard work week, this pattern indicates that weekends are likely to have much higher network participation than weekdays.

For the purposes of model adequacy checking, I will be using the full model, to demonstrate how this model can work with dependence terms in both the vertex and edge models. The Hamming error indicates that this model is the best fitting model. Exponential family models are guaranteed to conditionally reproduce the mean statistics for the sufficient statistics parameterized by the model, but in addition to these terms, we also examine several additional statistics. The statistics that we use for model checking are network size, density, diameter, and mean degree. These terms are described in figure 4.2.

In figure 4.2, the red line represents the observed statistic. The polygon represents the 95% simulation interval of this statistic, computed from 100 simulated draws of the network, with a polygon for each of the networks in the time series. If the red line falls within the polygon, then we consider the model to have adequately reproduced that network statistic.

These GOF terms indicate that the model is a reasonably well fitting model, as it does adequately reproduce the observed graph statistics in all cases. The graphs that are produced by the model are able to capture the seasonal pattern of network participation across the week. The simulation intervals for the edge processes tend to be a bit larger than the vertex process, but for most of the networks in the time series, the statistics we look at are within the 95% simulation intervals. This indicates that these models are able to produce networks that have the same structural character as the observed networking, providing some support for model adequacy.

4.6 Discussion

We present a new joint exponential family to understand and model joint vertex/edge dynamics. While we have demonstrated the model using an implementation of it using the R platform and the statistical package Statnet [Handcock et al., 2008b], the model is predicated on several assumptions which we not discussed in detail above. The main assumption is that the risk set of vertices that could enter the network is known. In other words, when fitting the model, every person who could be in the network is known. There are some cases where a model could be theorized in which the full risk set does not need to be known, such as the case where there is missing data in the observed networks (specifically regarding the nodes). This then becomes a missing data problem, and the appropriate methods could be applied. Recent work on this problem, in which there is an unknown infinite risk set has indicated that this is a plausible model family.

There are several interesting issues that are raised by this model, specifically around the model being inferentially separable but not separable under simulation. This choice in specifying our modeling framework allows for us to fit and evaluate models using an information criterion (e.g. AIC or BIC) more easily, as the models can be fit separable. However, this obscures that the vertex components of the model may often affect the edge components of the model, so when selecting a model only by AIC or BIC, additional checks for model adequacy should be performed to ensure that the model fits the data from a structural perspective.

Frameworks like HOPE (Held-Out Predictive Evaluation) may also prove useful for model selection [Yin et al., 2019]. Structural evaluations, like the one that we do in this chapter can also inform model selection and adequacy evaluation.

Finally, while we present a quorum of terms in this paper, the list of model terms presented here is not intended to be complete. This list of terms provides a starting point where some models can be fit, but in the coming years, additional terms may be presented that expand the framework to be better able to describe a variety of social phenomena. The ERGM framework has a huge collection of possible model terms already described and parameterized, but the vertex process does not have a fully developed list of model terms at this point.

4.7 Conclusion

Discrete time models prove to be an area where additional innovations may be had, especially in considering joint vertex and edge processes. We contribute to this literature through the joint vertex/edge model specified and implemented in this paper. We combine ideas from Generalized Location Systems [Butts, 2007] and ERGM [Morris et al., 2008] to produce a flexible framework for modeling and simulating from joint vertex and edge processes. This framework uses the ability of the ERGM modeling framework to model arbitrary dependence relationships in both the vertex and edge processes.

We encourage the continued development of these models, noting that the vertex terms that we discuss in this paper serve only as a starting point for modeling these networks. One important frontier is the availability of data for which these models can be used. It is our hope that the availability of more modeling options for these data types will encourage researchers to collect additional data for which there may be the joint vertex and edge processes discussed in this paper.

Chapter 5

Conclusion

In this thesis, I have explored several different ways in which space and geography affect social structures. The research contained in this dissertation highlights several important findings, which I have enumerated in no particular order here:

1. Understanding the local environments that people exist within (w.r.t. geography) significantly improves our understanding of how diseases that can only spread very locally move and disproportionately affect communities that may be near to each other.

2. It is important to understand the underlying processes at play in a social system when selecting or specifying a model. In the case of the COVID-19 research, the local spread process of the disease meant that many standard SIF models were not suitable for use in modeling COVID-19. For the third chapter, identifying the vertex dynamics process as one that should be modeled, rather than implicitly conditioned upon, opened the door to better modeling techniques for this type of dynamic network data.

3. Large scale processes often obscure a lot of local spatial heterogeneity. In the case of the COVID-19 research, while the pandemic affected all the cities we studied, the impact in different regions of these cities were very different. Our research highlighted that local network cohesion was a potential driver of the spatial heterogeneity. Additionally, when working with crosstab data, we had the higher order process (specified by the target distribution), and the local structural constraints (the 2-way marginals and integer constraints). While integrating the higher order process into our imputation

significantly improved prediction error, the local structure still imposed significant constraint on the possible three-way arrays.

4. While there are a variety of social processes that are affected by advantaged or disadvantaged social category membership, the superadditive relationships that are often described in literature exist alongside both additive and subadditive relationships. The crosstab work we did highlighted the importance of testing for non-linear relationships in the data, and the imporatnce of understanding how the level of spatial aggregation of data may affect those relationships.

Research is always an iterative process, and the work done here represents another step in a broader research agenda that I would like to pursue into the future. Drawing on the chapters I present and the findings above, I enumerate several different directions this research could go in the future.

1. The work on network models with demographic exchange produced a set of terms that act as a base for parameterizing the vertex component of the models. I would like to continue developing more terms, especially for the effects of environmental covariates, or situations in which the network may have a maximum capacity.

2. The dynamic network models should also be extended to specifically allow for the case in which a single population is spread across muliple network locations. This extension could begin work on a new type of migration model where nodes can move between spatial locations, or out of the network region.

3. The crosstab work was originally intended to be the first step in generating three-way population distributions at a fine level of geography in order to create new, inhomogenous Spatial Interaction Functions. With the imputation method implemented and validated, the development of new SIFs can now proceed.

Finally, I would like to address several of the challenges posed by the research in this dissertation:

1. One major challenge across the second and third chapters of the dissertation was the implementation of software for the imputation procedure and the modeling framework. While we set out to provide general purpose implementations, challenges during the research process meant that the final version of the software implementation was specific to the research contained within the chapters. In the future, it coud be valuable to retrofit these implementations for release as R packages.

2. For the third chapter on dynamic network models with demographic exchange, one limitation of this research is the requirement to have a known, constant risk set of vertices. For systems in which people enter or exit the risk population (e.g. migratory populations), this makes this model family far more difficult to use. If the model were to be extended to multiple network sites, the discrete time nature of the model would compound data collection efforts.

3. Some of the constraints on which types of crosstabs will be imputed using the framework specified by Bunea and Besag are more limiting than is preferable. Specifically, needing one of the dimensions of the array to have only 2 values constrains the types of data that can be imputed with our algorithm. Likewise, not having easy extensions to higher order arrays makes the future directions for this type of research more difficult to proceed with.

Bibliography

Alba, R. D., Logan, J. R., and Bellair, P. E. (1994). Living with crime: The implications of racial/ethnic differences in suburban location. *Social Forces*, 73(2):395–434.

Almquist, Z. W. (2020). Large-scale spatial network models: An application to modeling information diffusion through the homeless population of San Francisco. *Environment and Planning B: Urban Analytics and City Science*, 47(3):523–540.

Almquist, Z. W. and Butts, C. T. (2012). Point process models for household distributions within small areal units. *Demographic Research*, 26:593–632.

Almquist, Z. W. and Butts, C. T. (2013). Dynamic network logistic regression: A logistic choice analysis of inter- and intra-group blog citation dynamics in the 2004 US presidential election. *Political Analysis*, 21:430–448.

Almquist, Z. W. and Butts, C. T. (2014). Logistic Network Regression for Scalable Analysis of Networks with Joint Edge/Vertex Dynamics. *Sociological Methodology*, 44(1):273–321.

Almquist, Z. W. and Butts, C. T. (2015). Predicting regional self-identification from spatial network models. *Geographical analysis*, 47(1):50–72.

Alós-Ferrer, C., Hügelschäfer, S., and Li, J. (2016). Inertia and Decision Making. *Frontiers in Psychology*, 7:169.

Arabshahi, F., Huang, F., Anandkumar, A., Butts, C. T., and Fitzhugh, S. M. (2015). "are you going to the party: depends, who else is coming?" learning hidden group dynamics via conditional latent tree models. In *IEEE International Conference on Data Mining (ICDM)*.

Baker, W. E. (1984). The social structure of a national securities market. *American Journal of Sociology*, 89(4):755–811.

Bennett, R. J., Haining, R. P., and Griffith, D. A. (1984). The problem of missing data on spatial surfaces. *Annals of the Association of American Geographers*, 74(1):138–156.

Bhopal, S. S. and Bhopal, R. (2020). Sex Differential in COVID-19 Mortality Varies Markedly by Age. *Lancet (London, England)*.

Block, P., Hoffman, M., Raabe, I. J., Dowd, J. B., Rahal, C., Kashyap, R., and Mills, M. C. (2020). Social network-based distancing strategies to flatten the COVID 19 curve in a post-lockdown world. *Nature Human Behavior*.

Block, P., Hollway, J., Stadtfeld, C., Koskinen, J., and Snijders, T. (2022). Circular specifications and "predicting" with information from the future: Errors in the empirical saom–tergm comparison of leifeld & cranmer. *Network Science*, 10(1):3–14.

Boessen, A. and Hipp, J. R. (2015). Close-ups and the scale of ecology: Land uses and the geography of social context and crime. *Criminology*, 53:399:426.

Boessen, A., Hipp, J. R., Butts, C. T., Nagle, N. N., and Smith, E. J. (2017). Social fabric and fear of crime: Considering spatial location and time of day. *Social Networks*, 51:60–72.

Bossard, J. H. (1938). Ecological areas and marriage rates. *American Journal of Sociology*, 44(1):70–85.

Brockmann, D. and Helbing, D. (2013). The hidden geometry of complex, network-driven contagion phenomena. *Science*, 342(6164):1337–1342.

Bunea, F. and Besag, J. (2000). Mcmc in ixjxk contingency tables. *Monte Carlo Methods*, 26:25.

Butts, C. T. (2007). Models for generalized location systems. *Sociological Methodology*, 37(1):283–348.

Butts, C. T. (2008a). network: a package for managing relational data in R. *Journal of Statistical Software*, 24(2).

Butts, C. T. (2008b). Social network analysis with sna. *Journal of Statistical Software*, 24(6).

Butts, C. T. (2008c). Social Network Analysis with sna. *Journal of Statistical Software*, 24(6):1–51.

Butts, C. T. and Acton, R. M. (2011). Spatial modeling of social networks. *The Sage Handbook of GIS and Society Research. Thousand Oaks, CA: SAGE Publications*, pages 222–250.

Butts, C. T., Acton, R. M., Hipp, J. R., and Nagle, N. N. (2012a). Geographical variability and network structure. *Social Networks*, 34:82–100.

Butts, C. T., Acton, R. M., Hipp, J. R., and Nagle, N. N. (2012b). Geographical variability and network structure. *Social Networks*, 34(1):82–100.

Butts, C. T., Acton, R. M., and Marcum, C. S. (2012c). Interorganizational collaboration in the Hurricane Katrina response. *Journal of Social Structure*, 13.

Butts, C. T. and Almquist, Z. W. (2015). A Flexible Parameterization for Baseline Mean Degree in Multiple-Network ERGMs. *The Journal of Mathematical Sociology*, 39(3):163–167.

CDC (2020). CDC COVID-19 Pandemic Planning Scenarios. https://cdc.gov/coronavirus/2019-ncov/hcp/planning-scenarios.html. Accessed: 9/7/2020.

CDC (2021). COVID-19 Hospitalization and Death by Race/Ethnicity. https://www.cdc.gov/coronavirus/2019-ncov/covid-data/investigations-discovery/hospitalization-death-by-race-ethnicity.html. Accessed: 2/1/2021.

CDC COVID and Response Team (2020). Severe outcomes among patients with coronavirus disease 2019 (COVID-19)—United States, February 12–March 16, 2020. *MMWR Morb Mortal Wkly Rep*, 69(12):343–346.

Cinner, JE and Lau, JD and Bauman, AG and Feary, DA and Januchowski-Hartley, FA and Rojas, CA and Barnes, ML and Bergseth, BJ and Shum, E and Lahari, R and others (2019). Sixteen years of social and ecological dynamics reveal challenges and opportunities for adaptive management in sustaining the commons. *Proceedings of the National Academy of Sciences*, 116(52):26474–26483.

Clogg, C. C. and Eliason, S. R. (1987). Some common problems in log-linear analysis. *Sociological Methods & Research*, 16(1):8–44.

Cohen, M. L. and Di Zhang, X. (1988). The difficulty of improving statistical synthetic estimation.

Cranmer, S. J. and Desmarais, B. A. (2011). Inferential network analysis with exponential random graph models. *Political Analysis*, 19(1):66–86.

Crenshaw, K. (1990). Mapping the margins: Intersectionality, identity politics, and violence against women of color. *Stan. L. Rev.*, 43:1241.

Daraganova, G., Pattison, P., Koskinen, J., Mitchell, B., Bill, A., Watts, M., and Baum, S. (2012). Networks and geography: Modelling community network structures as the outcome of both spatial and network processes. *Social Networks*, 34(1):6 – 17.

Darroch, J. N. and Ratcliff, D. (1972). Generalized iterative scaling for log-linear models. *The annals of mathematical statistics*, pages 1470–1480.

Department of Homeland Security (2016). Homeland infrastructure foundation-level data.

Desmarais, B. and Cranmer, S. (2012). Statistical mechanics of networks: Estimation and uncertainty. *Physica A: Statistical Mechanics and its Applications*, 391(4):1865 – 1876.

Dowd, J. B., Andriano, L., Brazel, D. M., Rotondi, V., Block, P., Ding, X., Liu, Y., and Mills, M. C. (2020). Demographic science aids in understanding the spread and fatality rates of COVID-19. *Proceedings of the National Academy of Sciences*, 117(18):9696–9698.

Espuny-Pujol, F., Morrissey, K., and Williamson, P. (2018). A global optimisation approach to range-restricted survey calibration. *Statistics and Computing*, 28:427–439.

Facebook Connectivity Lab, Center for International Earth Science Information Network, CIESIN, and Columbia University (2016). High resolution settlement layer (HRSL). Source imagery for HRSL © 2016 DigitalGlobe, accessed May 31 2022.

Fan, C., Jiang, Y., and Mostafavi, A. (2020). Emergent Social Cohesion for Coping with Community Disruptions in Disasters. *Journal of the Royal Society Interface*, 17(164):20190778.

Ferguson, Neil and Laydon, Daniel and Nedjati Gilani, Gemma and Imai, Natsuko and Ainslie, Kylie and Baguelin, Marc and Bhatia, Sangeeta and Boonyasiri, Adhiratha and Cucunuba Perez, ZULMA and Cuomo-Dannenburg, Gina and others (2020). Impact of non-pharmaceutical interventions (NPIs) to reduce COVID19 mortality and healthcare demand. https://www.imperial.ac.uk/media/imperial-college/medicine/sph/ide/gida-fellowships/Imperial-College-COVID19-NPI-modelling-16-03-2020.pdf. Accessed: 11/18/2021.

Festinger, L., Schachter, S., and Back, K. (1950). *Social pressures in informal groups; a study of human factors in housing.* Harper.

Freeman, L. C., Freeman, S. C., and Michaelson, A. G. (1988). On human social intelligence. *Journal of Social and Biological Structures*, 11(4):415–425.

Geng, Y.-J., Wei, Z.-Y., Qian, H.-Y., Huang, J., Lodato, R., and Castriotta, R. J. (2020). Pathophysiological characteristics and therapeutic approaches for pulmonary injury and cardiovascular complications of coronavirus disease 2019. *Cardiovascular Pathology*, page 107228.

Geyer, C. J. and Thompson, E. A. (1992). Constrained monte carlo maximum likelihood for dependent data. *Journal of the Royal Statistical Society: Series B (Methodological)*, 54(3):657–683.

Goodreau, S. M., Kitts, J. A., and Morris, M. (2009). Birds of a feather, or friend of a friend?: Using exponential random graph models to investigate adolescent social networks. *Demography*, 46(1):103–125.

Graham, P., Young, J., and Penny, R. (2009). Multiply imputed synthetic data: Evaluation of hierarchical bayesian imputation models. *Journal of Official Statistics*, 25(2):245.

Handcock, M. S., Hunter, D. R., Butts, C. T., Goodreau, S. M., and Morris, M. (2008a). statnet: Software tools for the representation, visualization, analysis and simulation of network data. *Journal of Statistical Software*, 24(1):1–11.

Handcock, M. S., Hunter, D. R., Butts, C. T., Goodreau, S. M., and Morris, M. (2008b). statnet: Software Tools for the Representation, Visualization, Analysis and Simulation of Network Data. *Journal of Statistical Software*, 24(1):1548.

Hanneke, S. and Xing, E. P. (2007). Discrete temporal models of social networks. In Airoldi, E. M., an Stephen E. Fienberg, D. M. B., Goldenberg, A., Xing, E. P., and Zheng, A. X., editors, *Statistical Network Analysis: Models, Issues, and New Directions: ICML 2006 Workshop on Statistical Network Analysis, Pittsburgh, PA, USA, June 29, 2006, Revised Selected Papers*, volume 4503 of *Lecture Notes in Computer Science*, pages 115–125. Springer-Verlag.

Havers, F. P., Reed, C., Lim, T., Montgomery, J. M., Klena, J. D., Hall, A. J., Fry, A. M., Cannon, D. L., Chiang, C.-F., Gibbons, A., et al. (2020). Seroprevalence of antibodies to SARS-CoV-2 in 10 sites in the United States, March 23-May 12, 2020. *JAMA Internal Medicine*.

Hipp, J. R., Butts, C. T., Acton, R., Nagle, N. N., and Boessen, A. (2013). Extrapolative simulation of neighborhood networks based on population spatial distribution: Do they predict crime? *Social Networks*, 35(4):614–625.

Holland, P. W. and Leinhardt, S. (1981). An Exponential Family of Probability Distributions for Directed Graphs. *Journal of the American Statistical Association*, 76(373):33–50.

Hummel, R. M., Hunter, D. R., and Handcock, M. S. (2012). Improving simulation-based algorithms for fitting ergms. *Journal of Computational and Graphical Statistics*, 21(4):920–939.

Jackson, M. L., Hart, G. R., McCulloch, D. J., Adler, A., Brandstetter, E., Fay, K., Han, P., Lacombe, K., Lee, J., Sibley, T., et al. (2020). Effects of weather-related social distancing on city-scale transmission of respiratory viruses. *medRxiv*.

Jaynes, E. T. (1982). On the rationale of maximum-entropy methods. *Proceedings of the IEEE*, 70(9):939–952.

Jiang, F., Deng, L., Zhang, L., Cai, Y., Cheung, C. W., and Xia, Z. (2020). Review of the clinical characteristics of coronavirus disease 2019 (COVID-19). *Journal of General Internal Medicine*, pages 1–5.

Kadi, N. and Khelfaoui, M. (2020). Population density, a factor in the spread of COVID-19 in Algeria: Statistical study. *Bulletin of the National Research Centre*, 44(1):1–7.

Khalili, M., Karamouzian, M., Nasiri, N., Javadi, S., Mirzazadeh, A., and Sharifi, H. (2020). Epidemiological Characteristics of COVID-19: a Systematic Review and Meta-analysis. *Epidemiology & Infection*, 148.

Khubba, S., Heim, K., and Hong, J. (2022). National Census coverage estimates for people in the United States by demographic characteristics: 2020 post-enumeration survey estimation report. US Census Report PES20-G-01.

King, G., Rosen, O., and Tanner, M. A. (1999). Binomial-beta hierarchical models for ecological inference. *Sociological Methods & Research*, 28(1):61–90.

Krackhardt, D. and Porter, L. W. (1986). The snowball effect: Turnover embedded in communication networks. *Journal of Applied Psychology*, 71(1):50–55.

Krivitsky, P. N. (2012). Exponential-family random graph models for valued networks. *Electronic Journal of Statistics*, 6:1100–1128.

Krivitsky, P. N. and Butts, C. T. (2017). Exponential-family random graph models for rank-order relational data. *Sociological Methodology*, 47:68–112.

Krivitsky, P. N. and Goodreau, S. M. (2019). STERGM-Separable Temporal ERGMs for modeling discrete relational dynamics with statnet. (Tech. Rep.). Retrieved from https://cran.microsoft.com/snapshot/2016-06-21/web/packages/tergm/vignettes/STERGM.pdf.

Krivitsky, P. N. and Handcock, M. S. (2014). A separable model for dynamic networks. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(1):29–46.

Krivitsky, P. N., Handcock, M. S., and Morris, M. (2011). Adjusting for network size and composition effects in exponential-family random graph models. *Statistical Methodology*, 8(4):319–339.

Krivitsky, P. N. and Morris, M. (2017). Inference for social network models from egocentrically sampled data, with application to understanding persistent racial disparities in HIV prevalence in the US. *The Annals of Applied Statistics*, 11(1):427–455.

Kubrin, C. E. and Hipp, J. R. (2016). Do fringe banks create fringe neighborhoods? Examining the spatial relationship between fringe banking and neighborhood crime rates. *Justice Quarterly*, 33(5):755–784.

Kubrin, C. E., Hipp, J. R., and Kim, Y.-A. (2018). Different than the sum of its parts: Examining the unique impacts of immigrant groups on neighborhood crime rates. *Journal of Quantitative Criminology*, 34(1):1–36.

Lauer, S. A., Grantz, K. H., Bi, Q., Jones, F. K., Zheng, Q., Meredith, H. R., Azman, A. S., Reich, N. G., and Lessler, J. (2020). The incubation period of coronavirus disease 2019 (COVID-19) from publicly reported confirmed cases: estimation and application. *Annals of Internal Medicine*.

Lee, F. and Butts, C. T. (2020). Incorporating structural stigma into network analysis. *Social Networks*, 63:91–99.

Lee, S. (2018). Markov chain monte carlo and exact conditional tests with three-way contingency tables. Technical report, Naval Postgraduate School.

Lerner, J., Indlekofer, N., Nick, B., and Brandes, U. (2013). Conditional independence in dynamic networks. *Journal of Mathematical Psychology*, 57(6):275 – 283. Social Networks.

Leskovec, J., Kleinberg, J., and Faloutsos, C. (2007). Graph evolution: Densification and shrinking diameters. *ACM Transactions on Knowledge Discovery from Data*, 1(1).

Li, L., Zhang, J., Liu, C., Zhang, H.-T., Wang, Y., and Wang, Z. (2019). Analysis of transmission dynamics for Zika virus on networks. *Applied Mathematics and Computation*, 347:566–577.

Li, Q., Guan, X., Wu, P., Wang, X., Zhou, L., Tong, Y., Ren, R., Leung, K. S., Lau, E. H., Wong, J. Y., et al. (2020a). Early transmission dynamics in Wuhan, China, of novel coronavirus–infected pneumonia. *New England Journal of Medicine*, 382(13):1199–1207.

Li, R., Pei, S., Chen, B., Song, Y., Zhang, T., Yang, W., and Shaman, J. (2020b). Substantial undocumented infection facilitates the rapid dissemination of novel coronavirus (SARS-CoV-2). *Science*, 368(6490):489–493.

Logan, J. R. and Stults, B. J. (1999). Racial differences in exposure to crime: The city and suburbs of Cleveland in 1990. *Criminology*, 37(2):251–276.

Luna, A., Zhang, L.-C., Whitworth, A., and Piller, K. (2015). Small area estimates of the population distribution by ethnic group in england: a proposal using structure preserving estimators. *Statistics in Transition new series*, 16(4):585–602.

Lusher, D., Koskinen, J., and Robins, G. (2012). *Exponential Random Graph Models for Social Networks: Theory, Methods, and Applications*. Cambridge University Press, Cambridge.

Magesh, S., John, D., Li, W. T., Li, Y., Mattingly-App, A., Jain, S., Chang, E. Y., and Ongkeko, W. M. (2021). Disparities in covid-19 outcomes by race, ethnicity, and socioeconomic status: a systematic-review and meta-analysis. *JAMA network open*, 4(11):e2134147–e2134147.

Mallik, A. and Almquist, Z. W. (2019). Stable Multiple Time Step Simulation/Prediction From Lagged Dynamic Network Regression Models. *Journal of Computational and Graphical Statistics*, 28(4):967–979.

Massey, D. S. and Denton, N. A. (1993). *American apartheid: Segregation and the making of the underclass*. Harvard University Press.

Massey, D. S., Rothwell, J., and Domina, T. (2009). The changing bases of segregation in the united states. *The Annals of the American Academy of Political and Social Science*, 626(1):74–90.

Massey, D. S. and Tannen, J. (2018). Suburbanization and segregation in the United States: 1970–2010. *Ethnic and racial studies*, 41(9):1594–1611.

McHugh, M. L. (2013). The chi-square test of independence. *Biochemia medica: Biochemia medica*, 23(2):143–149.

McNulty, T. L. (1999). The residential process and the ecological concentration of race, poverty and violent crime in New York City. *Sociological Focus*, 32(1):25–42.

Molina, I. and Rao, J. (2010). Small area estimation of poverty indicators. *Canadian Journal of Statistics*, 38(3):369–385.

Moody, James and Adams, Jimi and Morris, Martina (2017). Epidemic potential by sexual activity distributions. *Network Science*, 5(4):461–475.

Mooney, S. J., Bader, M. D., Lovasi, G. S., Neckerman, K. M., Rundle, A. G., and Teitler, J. O. (2020). Using universal kriging to improve neighborhood physical disorder measurement. *Sociological Methods & Research*, 49(4):1163–1185.

Moretti, A. and Whitworth, A. (2020). Development and evaluation of an optimal composite estimator in spatial microsimulation small area estimation. *Geographical Analysis*, 52(3):351–370.

Morris, M. (2004). *Network epidemiology: a handbook for survey design and data collection*. Oxford University Press.

Morris, M., Epstein, H., and Wawer, M. (2010). Timing is everything: International variations in historical sexual partnership concurrency and HIV prevalence. *PloS ONE*, 5(11):e14092.

Morris, M., Handcock, M. S., and Hunter, D. R. (2008). Specification of Exponential-Family Random Graph Models: Terms and Computational Aspects. *Journal of Statistical Software*, 24(4):1548.

Morrison, P. A. (1971). Demographic information for cities: A manual for estimating and projecting local population characteristics.

New York Times (2020). Coronavirus (COVID-19) data in the United States. https://github.com/nytimes/covid-19-data.

Onder, G., Rezza, G., and Brusaferro, S. (2020). Case-fatality rate and characteristics of patients dying in relation to COVID-19 in Italy. *JAMA*.

Pattison, P. E. and Robins, G. L. (2002). Neighborhood-based models for social networks. *Sociological Methodology*, 32:301–337.

Pfeffermann, D. and Correa, S. (2012). Empirical bootstrap bias correction and estimation of prediction mean square error in small area estimation. *Biometrika*, 99(2):457–472.

Pfeffermann, D. et al. (2013). New important developments in small area estimation. *Statistical Science*, 28(1):40–68.

Pujari, B. S. and Shekatkar, S. M. (2020). Multi-city modeling of epidemics using spatial networks: Application to 2019-nCov (COVID-19) coronavirus in India. *medRxiv*.

R Core Team (2013). R: A Language and Environment for Statistical Computing.

R Core Team (2020). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rashed, E. A., Kodera, S., Gomez-Tames, J., and Hirata, A. (2020). Influence of absolute humidity, temperature and population density on COVID-19 spread and decay durations: Multi-prefecture study in Japan. *International Journal of Environmental Research and Public Health*, 17(15):5354.

Riley, S. (2007). Large-scale spatial-transmission models of infectious disease. *Science*, 316(5829):1298–1301.

Rose, A., McKee, J., Sims, K., Bright, E., Reith, A., , and Urban, M. (2021). Landscan global 2020 (data set).

Rosen, O., Jiang, W., King, G., and Tanner, M. A. (2001). Bayesian and frequentist inference for ecological inference: The r× c case. *Statistica Neerlandica*, 55(2):134–156.

Rubin, D. B. (1996). Multiple imputation after 18+ years. *Journal of the American statistical Association*, 91(434):473–489.

Sampson, R. J. and Sharkey, P. (2008). Neighborhood selection and the social reproduction of concentrated racial inequality. *Demography*, 45(1):1–29.

San Francisco Department of Public Health (2021a). COVID-19 Cases Over Time. https://data.sfgov.org/COVID-19/COVID-19-Cases-Over-Time/gyr2-k29z. Accessed: 10/07/2021.

San Francisco Department of Public Health (2021b). COVID-19 Cases Summarized by Race and Ethnicity. https://data.sfgov.org/COVID-19/COVID-19-Cases-Summarized-by-Race-and-Ethnicity/vqqm-nsqg. Accessed: 4/21/2021.

Saul, Z. M. M. and Filkov, V. (2007). Exploring biological network structure using exponential random graph models. *Bioinformatics*.

Schaefer, D. R., Bouchard, M., Young, J. T., and Kreager, D. A. (2017). Friends in locked places: an investigation of prison inmate network structure. *Social Networks*, 51:88–103.

Scholkopf, B. and Smola, A. J. (2001). *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press, Cambridge, MA.

Schweinberger, M., Krivitsky, P. N., Butts, C. T., and Stewart, J. (2020). Exponential-family models of random graphs: Inference in finite-, super-, and infinite-population scenarios. *Statistical Science*, 35(4):627–662.

Seidman, S. B. (1983). Network structure and minimum degree. *Social Networks*, 5:269–287.

Smith, E. J., Marcum, C. S., Boessen, A., Almquist, Z. W., Hipp, J. R., Nagle, N. N., and Butts, C. T. (2015). The relationship of age to personal network size, relational multiplexity, and proximity to alters in the Western United States. *Journals of Gerontology Series B: Psychological Sciences and Social Sciences*, 70(1):91–99.

Snijders, T. A. B., Pattison, P. E., Robins, G. L., and Handcock, M. S. (2006). New specifications for exponential random graph models. *Sociological Methodology*, 36:99–154.

Spiro, E. S., Almquist, Z. W., and Butts, C. T. (2016). The persistence of division: geography, institutions, and online friendship ties. *Socius*, 2:2378023116634340.

Steinberg, J. (1979). Synthetic estimates for small areas: Statistical workshop papers. *National Institute on Drug Abuse Research Monograph Series*, 24:282.

Stokes, E. K., Zambrano, L. D., Anderson, K. N., Marder, E. P., Raz, K. M., Felix, S. E. B., Tie, Y., and Fullerton, K. E. (2020). Coronavirus disease 2019 case surveillance—united states, january 22–may 30, 2020. *Morbidity and Mortality Weekly Report*, 69(24):759.

Thomas, L. J., Huang, P., Yin, F., Luo, X. I., Almquist, Z. W., Hipp, J. R., and Butts, C. T. (2020). Spatial heterogeneity can lead to substantial local variations in COVID-19 timing and severity. *Proceedings of the National Academy of Sciences*, 117(39):24180–24187.

Thomas, L. J., Huang, P., Yin, F., Xu, J., Almquist, Z. W., Hipp, J. R., and Butts, C. T. (2022a). Geographical patterns of social cohesion drive disparities in early covid infection hazard. *Proceedings of the National Academy of Sciences*, 119(12):e2121675119.

Thomas, L. J., Huang, P., Yin, F., Xu, J., Almquist, Z. W., Hipp, J. R., and Butts, C. T. (2022b). Geographical patterns of social cohesion drive disparities in early COVID infection hazard. *Proceedings of the National Academy of Sciences*, 119(12):e2121675119.

Tillman, B., Markopoulou, A., Butts, C. T., and Gjoka, M. (2019). 2K+ graph construction framework: Targeting joint degree matrix and beyond. *IEEE/ACM Transactions on Networking*, 27(2):591–606.

Townshend, I., Awosoga, O., Kulig, J., and Fan, H. (2015). Social Cohesion and Resilience Across Communities that have Experienced a Disaster. *Natural Hazards*, 76(2):913–938.

Verhagen, M. D., Brazel, D. M., Dowd, J. B., Kashnitsky, I., and Mills, M. (2020). Mapping hospital demand: demographics, spatial variation, and the risk of "hospital deserts" during COVID-19 in England and Wales. *OSF Preprints*.

Verity, R., Okell, L. C., Dorigatti, I., Winskill, P., Whittaker, C., Imai, N., Cuomo-Dannenburg, G., Thompson, H., Walker, P. G., Fu, H., et al. (2020). Estimates of the severity of coronavirus disease 2019: a model-based analysis. *The Lancet Infectious Diseases*.

Vermunt, J. K., Van Ginkel, J. R., Van der Ark, L. A., and Sijtsma, K. (2008). 9. Multiple imputation of incomplete categorical data using latent class analysis. *Sociological Methodology*, 38(1):369–397.

Viboud, C., Bjørnstad, O. N., Smith, D. L., Simonsen, L., Miller, M. A., and Grenfell, B. T. (2006). Synchrony, waves, and spatial hierarchies in the spread of influenza. *Science*, 312(5772):447–451.

Voinsky, I., Baristaite, G., and Gurwitz, D. (2020). Effects of Age and Sex on Recovery from COVID-19: Analysis of 5769 Israeli Patients. *Journal of Infection*, 81(2):e102–e103.

Wang, Q., Phillips, N. E., Small, M. L., and Sampson, R. J. (2018). Urban mobility and neighborhood isolation in America's 50 largest cities. *Proceedings of the National Academy of Sciences*, 115(30):7735–7740.

Wong, F. and Collins, J. J. (2020). Evidence that Coronavirus Superspreading is Fat-Tailed. *Proceedings of the National Academy of Sciences*, 117(47):29416–29418.

World Health Organization (2020a). Coronavirus disease 2019 (COVID-19): situation report, 43.

World Health Organization (2020b). *Media Statement: Knowing the risks for COVID-19*.

World Health Organization (2023). Who coronavirus (covid-19) dashboard.

Wu, J. T., Leung, K., and Leung, G. M. (2020). Nowcasting and forecasting the potential domestic and international spread of the 2019-nCoV outbreak originating in Wuhan, China: a modelling study. *The Lancet*, 395(10225):689–697.

Yaveroğlu, O. N., Fitzhugh, S. M., Kurant, M., Markopoulou, A., Butts, C. T., and Pržulj, N. (2015). `ergm.graphlets`: A package for ERG modeling based on graphlet statistics. *Journal of Statistical Software*, 65(12).

Yin, F., Phillips, N. E., and Butts, C. T. (2019). Selection of exponential-family random graph models via held-out predictive evaluation (hope). *arXiv preprint arXiv:1908.05873*.

Yu, Y., Grazioli, G., Unhelkar, M. H., Martin, R. W., and Butts, C. T. (2020). Network hamiltonian models reveal pathways to amyloid fibril formation. *Scientific Reports*, 10(1):1–11.

Zhang, Y., Jiang, B., Yuan, J., and Tao, Y. (2020). The impact of social distancing and epicenter lockdown on the COVID-19 epidemic in mainland China: A data-driven SEIQR model study. *medRxiv*.

Zipf, G. K. (2016). *Human behavior and the principle of least effort: An introduction to human ecology*. Ravenio Books.

Appendix A

Supplementary Information on Large City Simulations

A.0.1 Introduction

In this appendix, we go into more depth on Spatial Interaction Functions, Spatial Bernoulli Models, the setup and parameterization of our simulations, and the parameter estimates that were used for this paper. We also provide additional analyses regarding the role of local first-passage times as determinants of individuals' waiting times to infection, and the impact of alternative estimates of hospitalization rates on numbers of days at which hospitals would be expected to be at capacity.

A.0.2 Spatial Interaction Function

A Spatial Interaction Function (SIF) describes the marginal probability of a tie between any two nodes, given the distance between them. We denote the SIF by $\mathcal{F}(\mathcal{D}_{ij}, \theta)$, with $\mathcal{D}_{ij}$ being the distance between vertices i and j, and θ being the parameters for the function. Prior literature has found that spatial interaction functions for social networks like those of interest here tend to be of the power law or attenuated power law form Butts and Acton [2011]. Following this, we employ SIFs of the form $\mathcal{F}(\mathcal{D}_{ij}, \theta) = \frac{p_b}{(1+\alpha\mathcal{D}_{ij})^\gamma}$. Here, p_b represents the base tie probability, which can be thought of as the probability of a tie between two individuals residing at the same location. α is a scaling parameter that determines the phenomenological unit of distance for the decay in tie probability, and γ is the parameter that determines the weight of the tail (higher values imply fewer long-range ties, *ceteris paribus*).

We employ two SIFs in this paper, using models for social interactions and face-to-face interactions employed in prior studies Butts et al. [2012b], Hipp et al. [2013]. The social interaction SIF declines with a γ of 2.788, while the face-to-face SIF declines with γ of 6.437. The parameters for the social interaction SIF are $p_b = 0.533$, $\alpha = 0.032$, $\gamma = 2.788$, and the parameters for the face-to-face SIF are $p_b = 0.859$,$\alpha = 0.035$, $\gamma = 6.437$ Hipp et al. [2013].

A.0.3 Spatial Bernoulli Models

The Bernoulli Network Models are a class of random graph models in which each edge occurs as a Bernoulli trial, possibly with a distinct probability of occurrence. In a spatial Bernoulli graph, tie probabilities are determined by a Spatial Interaction Function, applied to the pairwise distances between individuals within some space (here, geographically determined using Census data). Spatial Bernoulli models are highly scalable due to the conditional independence of edges, but allow for extremely complex structure due to the heterogeneity in edge probabilities induced by the SIF; likewise, they naturally produce properties such as local cohesion and degree heterogeneity observed in many types of social networks Butts et al. [2012b]. Formally, we can specify a Spatial Bernoulli Model by $\Pr(Y_{ij} = 1) = \mathcal{F}(\mathcal{D}_{ij}, \theta)$, where Y_{ij} is a dichotomous indicator for the presence of the i, j edge, and $\mathcal{F}(\mathcal{D}_{ij}, \theta)$ is a Spatial Interaction Function taking as inputs the i, j distance $\mathcal{D}_{ij}$ and parameters θ.

A.0.4 Network Simulations

To simulate diffusion of COVID-19, we require a contact network. Here, we employ the above-described spatial Bernoulli graphs, with node locations for each of our 19 study locations drawn based on block-level Census data (including clustering within households, an important factor in disease diffusion). We follow the protocols described in Almquist and Butts [2012], Butts et al. [2012b] to generate node positions, specifically using the quasirandom (Halton) placement algorithm. Node placement begins with the households in each census block, using Census 2010 data with regions defined per Hipp et al. [2013]. The quasirandom placement algorithm uses a Halton sequence to place households in space within the areal unit in which they reside. If any two households are placed within a critical radius of each other, then the algorithm "stacks" the households on top of each other by introducing artificial elevation (simulating e.g. a multistory apartment building). Once all households are placed, individuals within households are placed at jittered locations about the household centroid. (Individuals not otherwise attached to households are treated as households of size 1.)

Given an assignment of individuals to spatial locations, we simulate spatial Bernoulli graphs using the models specified above. We generate two networks for each city, one with the social interaction SIF, and the other with the face-to-face interaction SIF. To form a network of potential high-risk contacts, we then merge these networks (which share the same node set) by taking their union, leading to a network in which two individuals are tied if they either have an ongoing social relationship or would be likely to have extensive face-to-face interactions for other reasons (e.g., interacting with neighbors). This process is performed for each city in our sample.

List of Cities

Table S1 lists the cities that we use for our simulations. These data are drawn from Hipp et al. [2013], with population data updated to reflect the most recent (2010) decennial Census.

A.0.5 Disease Simulations

We conduct a series of simulations to examine the spread of COVID-19 across city-sized networks. These simulations use a simple continuous-time network diffusion process, the general description of which are described in the main text. The input for the diffusion simulation is a network and a vector of initial disease states (*susceptible*, *latent* (infected but not yet infectious), *infectious*, *recovered*, and *deceased*), and the output is detailed history of the diffusion process up to the point at which a steady state is obtained (i.e.,

Table A.1: List of study communities.

	City/County
1	Buffalo
2	Baltimore
3	Cincinnati
4	Cleveland
5	Denver
6	Indianapolis
7	Miami
8	Milwaukee
9	Nashville
10	Pittsburgh
11	Rochester
12	Sacramento
13	Salinas
14	San Diego City
15	Seattle
16	St. Petersburg
17	Tampa
18	Tuscon
19	Washington DC

no infectious individuals remain). Infection occurs via the network, with currently infectious individuals infecting susceptible alters as Poisson events with a fixed rate. The transitions between latent and infectious, and infectious and either recovery or mortality are governed by gamma distributions estimated from epidemiological data. Table A.2 shows the estimated shape and scale parameters for the gamma distributions employed here. The parameters for waiting time to infectiousness are directly available in the Appendix of Lauer et al. [2020], while those for the recovery and death are estimated by matching the mean and standard deviation of durations reported in the literature Verity et al. [2020]. Selection into death versus recovery was made via a Bernoulli trial drawn at time of infection (thereby determining which waiting time distribution was used), with the estimated mortality probability being 0.0138 using the case fatality rate adjusted for under-ascertainment reported in Verity et al. [2020]. Under these parameters, the median time to infectiousness is 5.2 days, with 95% of cases falling between 2 and 10.1 days; once infectious, respective median times to recovery and death are 25.1 days (95% range 9.6 to 52.2) and 16.9 days (95% range 5.5 to 38.3). We note in particular that these distributions incorporate the clinically observed skewness in recovery times, with many cases resolving in less than two weeks but a non-trivial fraction persisting for six weeks or longer.

Table A.2: Shape and Scale parameters for Gamma distributions for durations (unit: day).

	Death	Recovery	Infectious
Shape	4.566	5.834	5.807
Rate	0.251	0.219	1.055
Scale (i.e., 1/Rate)	3.984	4.566	0.948

A.0.6 Infection Rate Parameter Estimation

To determine the infection rate (the only free parameter for the models used in our simulations), we simulate the diffusion of virus in Seattle and fit it to the over-time death rate of the King County, WA before the first shelter-in-place order went into effect on March 23, 2020. We limit our data to this time period because our simulation employs a no-mitigation scenario. A grid search strategy was employed to determine the expected

days to transmission (which is the inverse of infection rate), and the number of days between the existence of the first infected cases and the first confirmed cases (aka the time lag, a nuisance parameter that is relevant only for estimation of the infection rate). The time lag is treated as an integer and the expected days to transmission as a continuous variable. For each lag/rate pair, we randomly take 5 draws from the expected infection waiting time distribution, add them to the lag time (i.e. the introduction of the true patient zero for the initial outbreak), and simulate 50 realizations of the diffusion process (redrawing the network each time). The diffusion rate parameter was selected based on minimizing the mean squared error between the simulated death rate and the observed death rate over the selected period. The first round of grid-search divided the expected days of search into 100 intervals, from (0,1) to (99,100), with days of lag ranging from 1 to 100 days. The second round of grid-search, based on the performance of the first round, divided the expected days of search into 240 intervals, from (40.00,40.25) to (99.75,100.00), with days of lag ranging from 1 to 60 days. The grid-search suggests that the expected days to transmission is 82.875 (82.75,83.00) days (Fig A.1); that is, in a hypothetical scenario in which a single infective ego remained indefinitely in the infective state, and a single alter remained otherwise susceptible, the average waiting time for ego to infect alter would be approximately 80 days. While this may at first blush appear to be a long delay, it should be borne in mind that this embodies the reality that no individual is likely to infect any *given* alter within a short period (since, indeed, ego and alter may not happen to interact within a narrow window). With many alters, however, the chance of passing on the disease is quite high. Likewise, we note that the thought experiment above should not be taken to imply that actors remain infectious for such an extended period of time; per the above-cited epidemiological data, individuals typically remain infectious for roughly 25.1 days (95% range 9.6-52.2). When both delay times are considered, the net probability of infecting any given alter prior to recovering is approximately 27%. We further calculated the corresponding basic reproductive number (R_0), which is the product of the probability of infection (27%) and the mean degree of the networks (10). The corresponding basic reproductive number in the diffusion simulation model is 2.7, compatible with estimates of R_0 at the pre-mitigation stage in other literature Li et al. [2020b,a], Wu et al. [2020].

A.0.7 Timing and Shape of Infection Curves

While our simulations show substantial heterogeneity in infection patterns across tracts, it may be hypothesized that this pattern is driven by a characteristic pattern of infection *within tracts*, with the heterogeneity being driven by the differences in *arrival time* of the infection to different tracts. (Such a pattern has been hypothesized by e.g. Brockmann and Helbing [2013] at the national level to explain heterogeneity at global scales.) To examine this possibility, we ran several additional analyses. First, to assess whether diffusion patterns within tracts follow a *universal curve*, we compare standardized within-tract prevalence curves across the entire sample of cities. To standardize the infection curves, we apply the following transformations. First, all selected infection prevalence curves (i.e., active cases as a function of time) were translated to the origin, such that the first infection occurs at time 0. Next, we standardized the maximum prevalence by dividing the number of infections active on any day by the maximum prevalence for the tract. Finally, we standardized the time scale of the prevalence curve so that the mean prevalence time (i.e., the centroid of the temporal distribution) occurs at unit time. The resulting standardized curves reflect the pure shape of the infection trajectory; if infection follows a universal pattern at local scales (up to an affine transformation), then the standardized curves should be approximately identical.

To assess the standardized curves, we chose a stratified sample of tracts from across the sample, selecting 5 tracts at random from each city. To avoid artifacts from tracts with insufficient infections, we exclude tracts with fewer than 20 infections during the time course of the simulation. Figure A.2 shows the resulting distribution of standardized prevalence curves. In line with the universality hypothesis, there is a dominant functional form approximately followed by a large fraction of tracts. However, we also observe a large fraction of tracts that deviate from this central pattern. Modes of deviation are idiosyncratic, and include multi-modality, substantial differences in skewness, and differences in "roughness" over time. Quantitatively, relatively large variations in standardized time to maximum prevalence are observed, with maxima for the majority of the sample falling between approximately 0.5 and 1.5 time-to-mean-prevalence units. These

observations suggest that, while there is a fairly *common* tract-level infection pattern, this pattern coexists with a wide range of other types of trajectories and is better thought of as a central tendency than a universal phenomenon.

Even if there is not a universal tract-level diffusion curve, it could still be the case that the majority of heterogeneity in infection times could be explained by the time taken for the infection to reach each tract: in particular, if the city-level diffusion process reflects a union of very small outbreaks with the waiting time for the infection to "jump" from one area to another being large compared to the time needed for the local outbreaks to reach all susceptibles within the local area, then the primary determinant of individual infection time would be the time taken for the infection to reach the individual's local region. We examine this hypothesis in figure A.3. The left-most boxplot of figure A.3 shows the distribution of times to infection for all individuals in our sample, timed from the onset of the first infection. The right-most boxplot shows the corresponding waiting times for the same individuals, *net of the first appearance of the infection in their tracts.* If infection time were determined primarily by the time it takes for the infection to reach an individual's local area, we would see a substantially compressed distribution relative to the total waiting time; on the contrary, the two are nearly the same, falsifying the conjecture that diffusion to the local area is the driver of waiting time heterogeneity.

The reason for this lack of compression can be seen in the middle boxplot of figure A.3, which shows the distribution of waiting times for the first case in each tract. We can see that, while individual infection times vary markedly, first passage times to tracts are both short and highly compressed: once introduced, the infection diffuses rapidly to nearly all tracts. Further, we can see that nearly all tracts are reached long before the majority of individuals are infected – thus, diffusion to local areas is not the primary limiting factor determining individual waiting times. Instead, the relative permeability of those areas to diffusion (which may depend upon factors such as local population density and the presence of barriers to interaction) appears to play a much greater role in governing the distribution of infection times.

A.0.8 Robustness of Spatial Heterogeneity on Hospital Load

Currently, considerable uncertainty exists regarding the fraction of SARS-CoV-2 infections leading to hospitalization; as such, it is useful to verify that the overall patterns of spatial heterogeneity seen here are robust to the hospitalization rate. On the high end, the World Health Organization and the CDC have estimated the rate to be 20% and 20-30%, respectively, in the absence of mitigation CDC COVID and Response Team [2020], World Health Organization [2020b]. Arguably, these estimates are inflated by under-reporting of asymptomatic or mildly symptomatic infections, particularly given the poor state of testing in the early stage of the pandemic. At the opposite extreme, one recent serology study contended that less than 10% of infections were reported in USA Havers et al. [2020], implying a potential hospitalization rate in the neighborhood of 2%. This study, too, faces problems with selection, as population prevalence was based on a convenience sample from patients seeking health care during the pandemic (and who are hence disproportionately likely to be infected); the true extent of under-reporting is thus likely to be smaller than this estimate (and the hospitalization rate correspondingly higher). Taking these two estimates as upper and lower bounds on the likely rate of COVID-19 hospitalizations, we replicated our analysis on hospital load using 20% and 2% (i.e., 10% of 20%) as respective probabilities of hospitalization per infection.

Fig. A.4 shows the respective marginal distributions of hospital overload periods with 20% and 2% hospitalization rates. Despite the majority of hospitals running at capacity for a relatively short period of time, a sizable fraction of hospitals experience overloads for very long periods. The persistence of this pattern over a full order of magnitude variation in hospitalization rates demonstrates the potential of unmitigated COVID-19 infections to severely strain local resources even under fairly optimistic scenarios, and suggests that substantial inequalities in healthcare service demand are robust to detailed hospitalization rates. However, we also note that the current level of uncertainty in hospitalization rates serves as a significant obstacle to quantitative prediction of the spatial distribution of hospital load for planning or response purposes.

Particularly given the large gap between anticipated load on "typical" units and those expected to be hit hardest, more refined rate estimates would seem to have the potential to inform important resource allocation decisions.

A.0.9 Code and Data Availability

Code and data needed to replicate the simulation and analysis for this paper can be found at:

Loring J. Thomas; Peng Huang; Fan Yin; Xiaoshuang Iris Luo; Zack W. Almquist; John R. Hipp; Carter T. Butts, 2020, "Replication Data for: Spatial Heterogeneity Can Lead to Substantial Local Variations in COVID-19 Timing and Severity," `https://doi.org/10.7910/DVN/B9XKSR`, Harvard Dataverse.

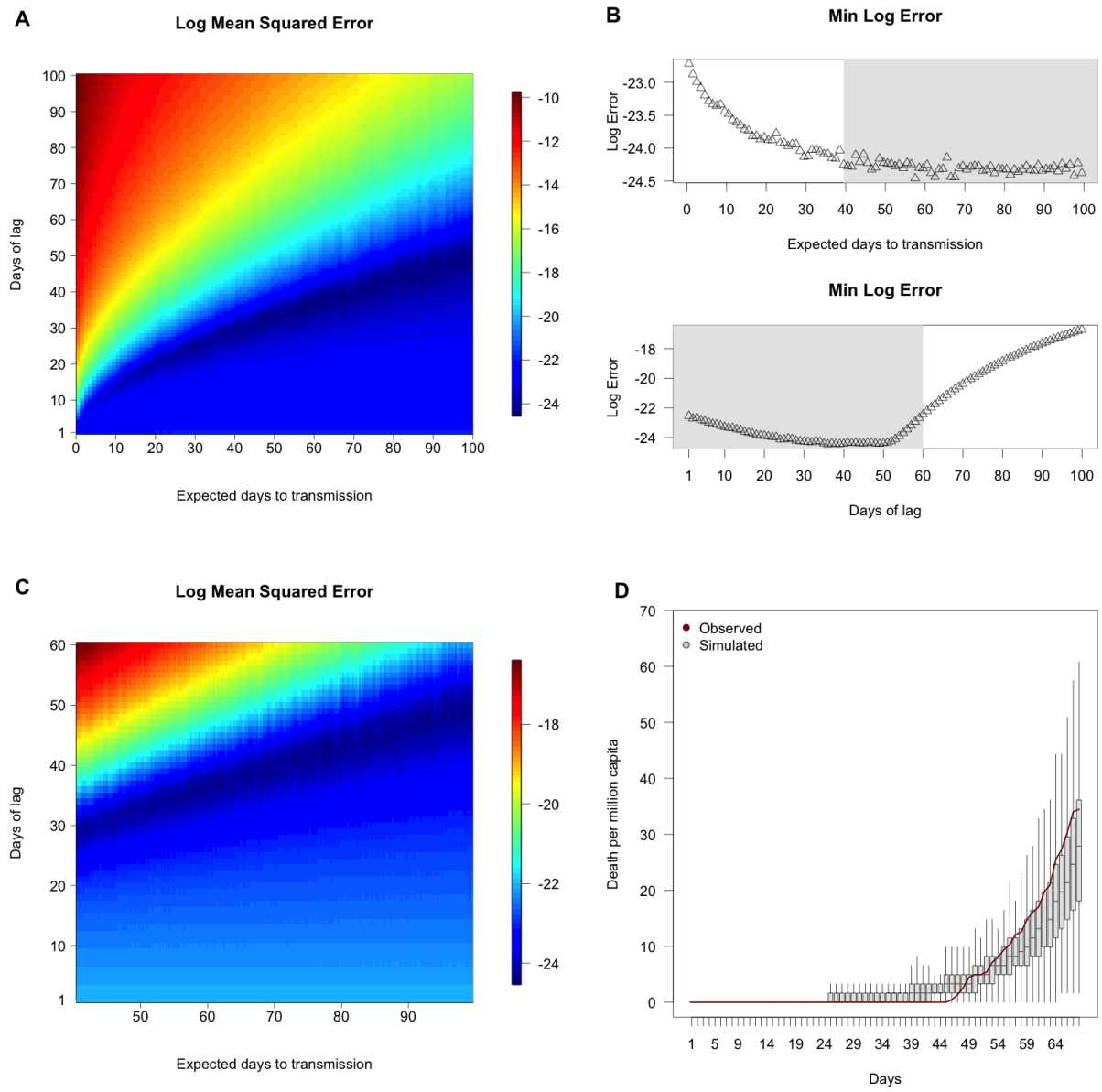


Figure A.1: **Grid-search for infection rate.** The mean squared error (in logarithm form) for simulations with combinations of days of lag and expected days to transmission, for the first round (**A**) and the second round (**C**). The cross-sectional analysis of the minimum log error for each days of lag and expected days to transmission (**B**) suggests the interval of both variables for the second round of search (areas in gray). Curves of death rate based on the best-fit parameter: 44 days of lag, 82.875 expected days to transmission (**D**).

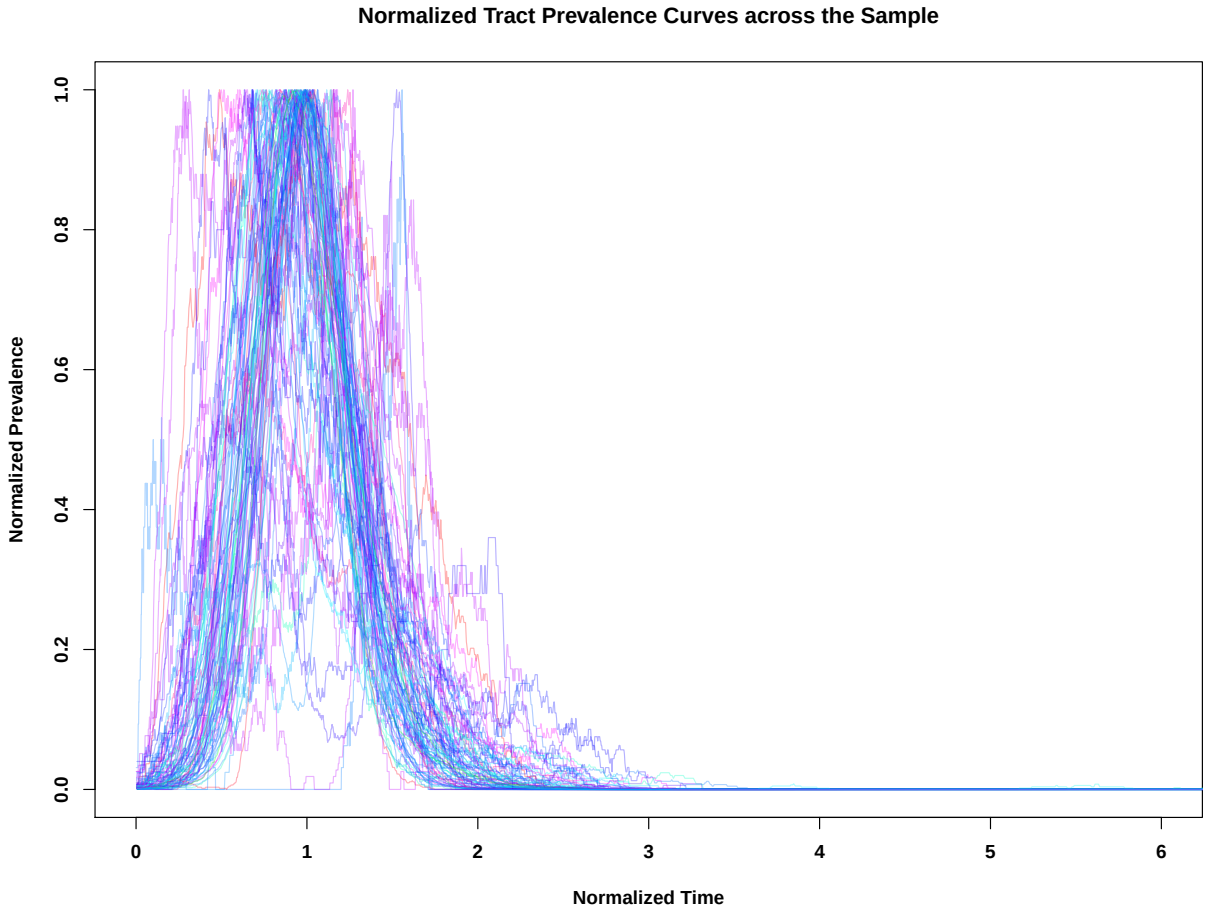


Figure A.2: Standardized infection curves at the tract level. Many tracts approximately follow the stereotypical "bell-shaped" pattern, but a large number of tracts deviate by being irregular, multimodal, and/or long-tailed. (Note: all trajectories scaled to unit maximum; apparent truncation actually reflects between-curve variation in time to maximum prevalence.)

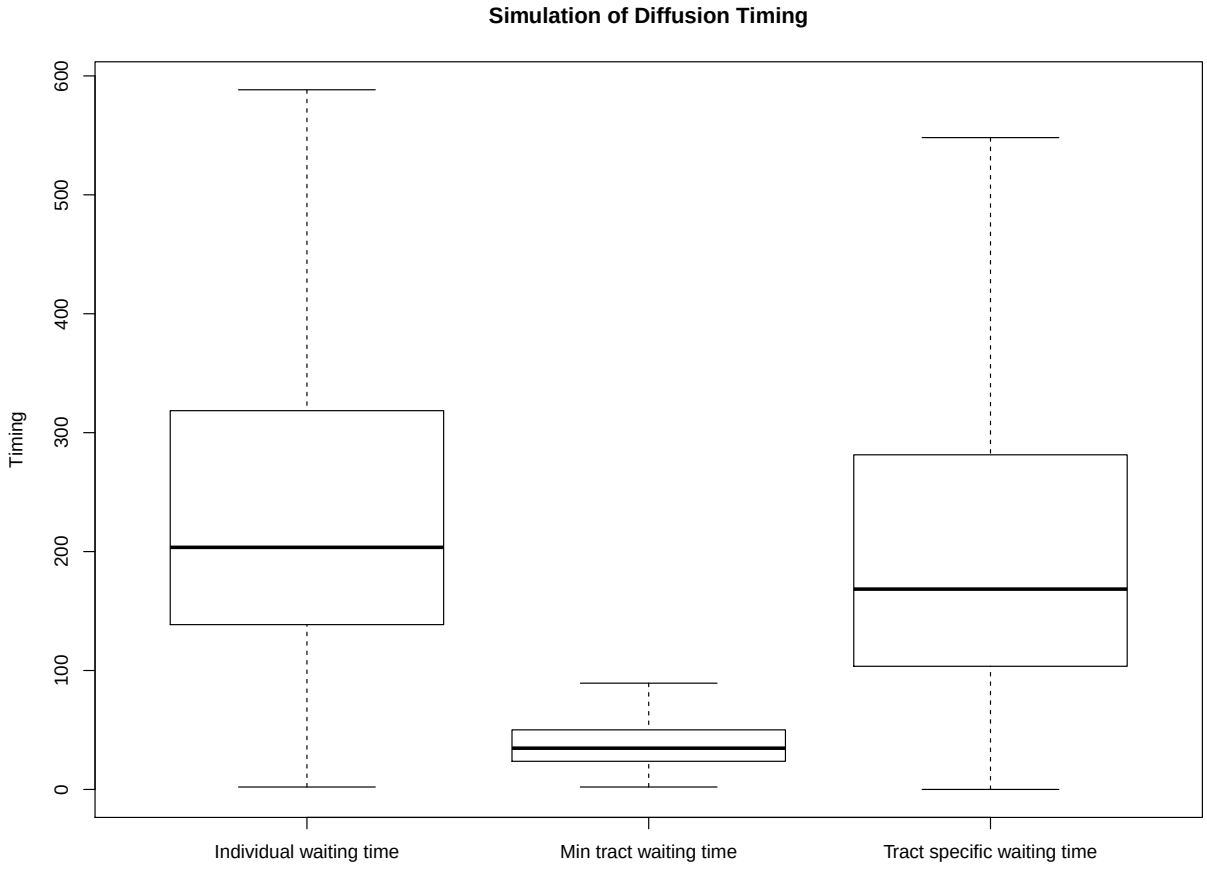


Figure A.3: Respective distributions of individual infection times (left), tract arrival times (middle), and tract specific infection times (right) for the tract sample. Individual infection times relative to arrival within tracts show little difference from infection times relative to the start of the larger outbreak, as tract arrival times are substantially shorter than the time needed to diffuse within tracts.

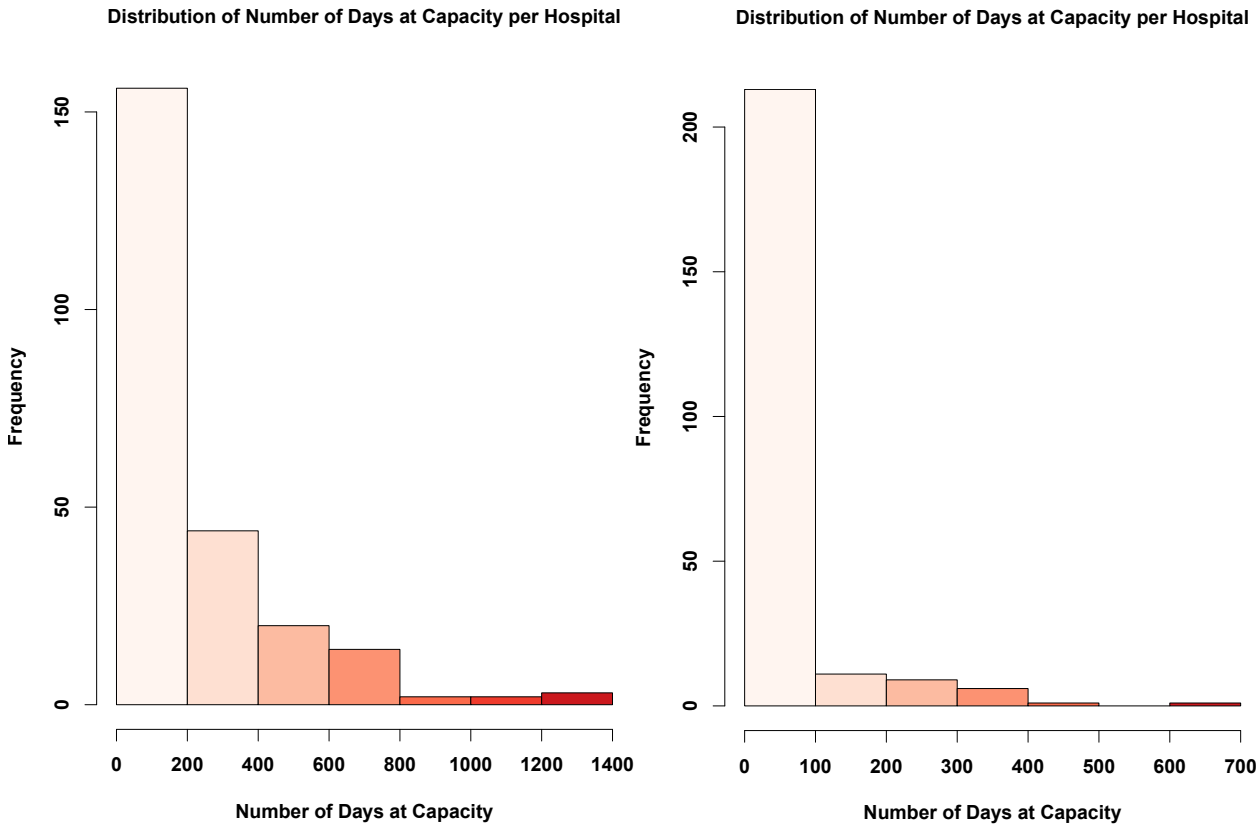


Figure A.4: (Left) The distribution of days that hospitals are at capacity, with a 20% hospitalization rate. (Right) The distribution of days that hospitals are at capacity with a 2% hospitalization rate. While higher hospitalization rates lead to higher levels of strain overall, extreme inequality in load persists in both scenarios.

Appendix B

Supplementary Information on San Fransicso Analysis

B.0.1 Introduction

In this appendix, we include additional information about the parameterization of the diffusion model, as well as the Cox Proportional-Hazards model. This section also provides more detail on the data that is used to generate the networks and estimate parameters.

B.0.2 Network and Demographic Data

We employ data from the 2010 U.S. Census to generate the population level social networks that underlie the analysis in this manuscript. Specifically, we use the smallest level of geography publicly available from the U.S. Census, known as the U.S. Census block level (approximately a city block in an urban setting). Each block contain basic demographic information, including household size.

To generate the network, we employ spatial network models that rely on a kernel function (the Spatial Interaction Functions, SIFs) to describe the presence of a social tie based on the distance between nodes; each node represents a single individual, and all simulations explicitly track the infection history of each individual in the population (as well as their infection paths). We employ the same network generation process used by Thomas et al. Thomas et al. [2020], which leverages the strategy of Almquist and Butts [2012], Butts et al. [2012a] of placing households within Census blocks using a low-discrepancy (Halton) sequence, followed by jittered placement of individual locations about the household center. To parameterize the model used in this manuscript, we need to first define the spatial network models (or spatial Bernoulli models) which depend on the SIF. The SIF describes the probability of a tie being present between any two entities, given the distance between those entities. We use the same SIFs as the ones used to generate networks in the large sample of cities, which employ a power law model of the form, $\mathcal{F}(\mathcal{D}_{ij}, \theta) = \frac{p_b}{(1+\alpha\mathcal{D}_{ij})^{\gamma}}$, where p_b describes the baseline probability of a tie existing, α is a scaling parameter describing the effect of a unit of distance, $\mathcal{D}_{ij}$ is the distance a dyad spans, and γ is a parameter describing the form of the tie probability decay. The simulation process employed uses two SIFs, based on prior literature to generate networks. The parameters for these SIFs are the same ones that are used for the large scale city simulations described in Appendix A.

Departing from prior work, we also leverage demographic information on U.S. Census blocks. These demo-

graphic covariates are race, ethnicity, age, and sex. These demographic covariates were assigned to nodes such that the three way distribution of race/sex/age and the two way distribution of race/ethnicity match the observed data at the block level. This allows a more fine-grained parameterization for simulation of the diffusion of COVID across social contact networks, based on demographic characteristics of each node (as detailed in the next section). We note that our procedure also leverages household size and thus represents the increased likelihood of being in a clique for individuals in such settings. This factor is one of the core factors that leads to COVID risk, as household spread of the disease is a primary avenue of spread.

We apply this technique to map social contact networks of San Francisco for three core reasons. (i) San Francisco is a city/county administrative unit – this is important because most data reported for the COVID-19 pandemic is at the county level in the U.S. and this allows us to analyze a complete city. (ii) San Francisco is a peninsula that is separated on three sides by water, reducing boundary effects from contacts outside the border of the city. (iii) The city/county of San Francisco published longitudinal data on infections by ethno-racial groups of the early pandemic San Francisco Department of Public Health [2021b]. The combination of good data management and reporting makes San Francisco unique, and when taken together with its status as a natural reporting unit (i.e. also being a county) it becomes an important unique case for studies such as the one conducted in this manuscript. We observe that future decisions by other municipalities to publish longitudinal data broken down by demographics would facilitate further studies of this kind.

B.0.3 Parameterization of Diffusion Model

To simulate the spread of COVID across a social contact network, we use a continuous time diffusion model defined by Thomas et al. [2020], and used for the large scale city simulations in this chapter. This diffusion model describes the way that individuals in the social network experience the disease and spread it to others. This diffusion model begins with the network structure and a vector of disease states for each node (individual). Disease states can be Susceptible (an individual who does not have the disease, but can get infected with it), Infected (the individual has been infected with the disease, but is not infectious), Infectious (the individual can spread the disease to others), Dead, or Recovered. At the beginning of the simulation, all nodes begin in the Susceptible state, with the exception of the seed infections. These nodes begin the simulation being infected with the disease. 25 individuals, randomly selected from the population, are the seed infections in each of the simulations.

Simulations are run until a steady state has been achieved, in which there are no more infected or infectious people, with everyone being in the Susceptible, Recovered, or Dead states. At this point, the diffusion model provides a detailed history for each node, describing the individual's final state in the simulation, as well as the times at which the node entered any given state. The disease spreads across the structure of the network, with connected nodes being able to transmit the disease across their social ties. Infection occurs as a Poisson event with a fixed rate, described by Thomas et al. [2020]. Only infectious nodes can infect susceptible social contacts; once an individual recovers or dies, they are no longer able to infect or be infected with COVID. When a Susceptible node is infected by an infectious alter, a Bernoulli trial is performed, determining whether a node becomes terminally or non-terminally infected. The rate of success (terminal infection) of the Bernoulli trial is given by P_d, a matrix sorted by age in the row and sex in the column (top to bottom row: 0-9, 10-19, 20-29, 30-39, 40-49, 50-59, 60-69, 70-79, and 80+; left to right column: female and male); for an individual with age category i and sex category j, the indicator for terminal infection thus arises as $T_{ij} \sim \text{Bern}(P_{dij})$. P_d, which is in essence a transformation of the Infection Fatality Ratio (IFR) broken down by age and sex, is calculated based on two pieces of information: the IFR for each age group Ferguson, Neil and Laydon, Daniel and Nedjati Gilani, Gemma and Imai, Natsuko and Ainslie, Kylie and Baguelin, Marc and Bhatia, Sangeeta and Boonyasiri, Adhiratha and Cucunuba Perez, ZULMA and Cuomo-Dannenburg, Gina and others [2020], and the sex ratio of death probability within each age group Bhopal and Bhopal [2020], assuming the probability of male and female getting infected is equal within each age group. P_d describes the set of Bernoulli parameters determining the likelihood of a fatal infection:

$$P_d = \begin{pmatrix} 0.000022 & 0.000018 \\ 0.000049 & 0.000049 \\ 0.000216 & 0.000384 \\ 0.000604 & 0.000996 \\ 0.001045 & 0.001955 \\ 0.003625 & 0.008375 \\ 0.012360 & 0.036410 \\ 0.030357 & 0.071643 \\ 0.070189 & 0.115811 \end{pmatrix}$$

The timing of transitions between different states is governed by a series of Gamma distributions. The waiting time from being infected to being infectious is governed by a Gamma distribution with shape 5.807 and scale 0.948, as estimated by Lauer et al. [2020]. For transition towards recovery or death, while prior work used homogeneous distributions, we break them down by demographics to more accurately account for variation across different populations. We estimate their parameters by matching the mean and standard deviation of waiting time for each group, using epidemiological data reported in Voinsky et al. [2020], Khalili et al. [2020], CDC [2020]. These method of moments estimators coincide with maximum likelihood estimators for the associated parameters, given that the Gamma distribution is a member of the exponential family. Specifically, the waiting time to death for a terminally infected individual in age category i is distributed as ${t^d}_i \sim \text{Gamma}(G_{di1}, G_{di2})$, where G_d is a parameter matrix whose columns contain shape and rate parameters, respectively, and rows indicate age category (top to bottom: 0-49, 50-64, and 65+). (Note that we do not vary the waiting time distribution by sex, as we are not aware of applicable time-to-mortality data from the early pandemic that supports age/sex decomposition.) Here, G_d is given as follows:

$$G_d = \begin{pmatrix} 3.744 & 0.251 \\ 3.568 & 0.233 \\ 2.881 & 0.223 \end{pmatrix}$$

The waiting time to recovery is broken down by both age and sex. For a male in age category i with a non-terminal infection, the waiting time to recovery is distributed as $t^r_{im} \sim \text{Gamma}(G^{rm}_{i1}, G^{rm}_{i2})$, where G^{rm} is a parameter matrix whose rows are ordered by age category (top to bottom: 0-19, 20-29, 30-39, 40-49, 50-59, 60+) and whose columns respectively contain shape and rate parameters. Here, G^{rm} is given as follows:

$$G^{rm} = \begin{pmatrix} 5.339 & 0.392 \\ 5.782 & 0.414 \\ 5.808 & 0.402 \\ 6.686 & 0.452 \\ 6.301 & 0.425 \\ 6.242 & 0.424 \end{pmatrix}$$

For a non-terminally infected female in age category i, the waiting time to recovery is similarly distributed as $t^r_{if} \sim \text{Gamma}(G^{rf}_{i1}, G^{rf}_{i2})$, where G^{rf} is a second parameter matrix whose rows are also ordered by age category (top to bottom: 0-19, 20-29, 30-39, 40-49, 50-59, 60+) and whose columns respectively contain shape and rate parameters. G^{rf} is as follows:

$$G^{rf} = \begin{pmatrix} 5.395 & 0.408 \\ 5.623 & 0.402 \\ 5.326 & 0.376 \\ 6.258 & 0.424 \\ 5.776 & 0.407 \\ 4.719 & 0.337 \end{pmatrix}$$

Since the diffusion process precedes the reporting of the first confirmed positive case, we performed a grid search to determine the length of the time lag between the appearance of "patient zero" in the city and the report of the first positive confirmed case (March 3, 2021 San Francisco Department of Public Health [2021a]). Our search was performed over an interval from a minimum of 1 and a maximum of 100 days. For each possible number, we regressed the number of infection case for each racial group in their observed time period using data from San Francisco Department of Public Health [2021b], on its counterparts in the simulation. The loss function is the summation of the mean squared errors (MSE) for all the linear regressions. We find that a 35 day lag minimizes the MSE, and this value is used here.

B.0.4 Simulation Details

Given the network and diffusion models described above, we run a series of simulations in which the population of San Francisco is seeded with randomly placed infectives 35 days prior to the first confirmed case report in San Francisco on March 3, 2021, and the infection process is followed until the end of our observation period (March 24, 2020, one week after demographic data becomes available for all four major racial/ethnic groups within the city). 35 individual-level contact networks were generated for San Francisco, using different simulated node locations for each realization. For each of these 35 simulated networks, we run 35 diffusion replicates, reseeding the seed infections for each simulation. This produces 1225 simulation replicates.These networks were produced with the R programming language, using the `sna` library Butts [2008c], R Core Team [2013]. For results reported about a single network realization in the main text, we average the infection time (or inverse infection time) for each diffusion replicate simulated in that network. The network being averaged across was selected as the network that most closely matches the average infection and susceptibility splits across all networks on March 24, 2020. For other metrics (such as the reported Cox model results), we average across the entire sample of networks. All figures from the main text utilize simulated data calibrated to observed data on infections and deaths.

The number of replications (independently simulated networks and diffusion simulations within network) was chosen based on a preliminary power analysis based on pilot simulations. Due to the diffusion simulation being bound to the structure of the social network, multiple network replicates were used to highlight trends in infection patterns across space. Likewise, given that the pandemic trajectories are dependent on the seed locations in the network, we randomized the seeds in each pandemic replicate to ensured that simulated trends were not due to idiosyncrasies in seed placement in the network structure. (The equality between the replication count and the inferred optimal lag time for the first infection is coincidental.)

B.0.5 Cox Proportional-Hazards Models

To assess the effects of local cohesion on infection hazards, we use Cox Proportional-Hazards models. Cox models control for (possibly time-varying) background hazards, allowing us to identify the impact of cohesion on infection hazard net of the overall progress of the outbreak. Because each simulated outbreak follows a distinct trajectory, we fit a single model to each simulated trajectory (with the baseline hazard, plus a single effect for core number). This model predicts the hazard of an uninfected individual getting infected with COVID-19, using the core number of a given node Seidman [1983] as a cohesion measure. The *core number* of a node - specifically, the highest k such that the node belongs to the kth degree core of the contact network - is a measure of local cohesion, with higher numbers indicating that the focal node is embedded in a more cohesive subgroup. In particular, nodes with core numbers of 0 are isolates, those of core 1 belong to trees or pendant trees, and those of core number 2 or higher belong to bicomponents (with higher numbers indicating higher levels of of cohesion). The core number is measured in units of ties, with a core number of k indicating that ego has at least k ties to alters who themselves have core numbers of at least k (and hence who have at least k ties to others with at least k ties to others in the core, recursively). We note that core number is not equivalent to degree: one can have arbitrarily high degree and still have a core number as

low as 1. The Cox model coefficient for core number thus indicates the extent to which nodes embedded in locally cohesive regions within the contact network are infected more or less rapidly (on average) than other nodes, controlling for the time-varying baseline infection hazard.

The form of the Cox used here is $h(t) = h_b(t)\exp(\beta X)$. Here, $h(t)$ represents the infection hazard, with $h_b(t)$ being the baseline hazard, X the core number, and β a coefficient expressing the increase in the log infection hazard per unit increase in core number. Here, we observed a mean β of 0.2615 over all simulations, implying an average risk enhancement of approximately 30% in infection hazard per unit increase in core number (as reflected in Fig.6C). As described in the main text, cohesion is a strong and consistent risk factor for early COVID infection, with nodes in high-order cores having a much higher infection risk than those in low-order cores.

B.0.6 Code and Data Availability

We have provided the code and data used for this project, including all parameters for the demographic models. This archive can be found at https://doi.org/10.7910/DVN/NT4KDH.