

THE INFLUENCE OF COLLECTIVE ANIMAL MOVEMENT ON POPULATION DYNAMICS

A Dissertation

Presented to the Faculty of the Graduate School
of Cornell University

in Partial Fulfillment of the Requirements for the Degree of
Doctor of Philosophy

by

Benjamin D. Dalziel

May 2014

© 2014 Benjamin D. Dalziel

ALL RIGHTS RESERVED

THE INFLUENCE OF COLLECTIVE ANIMAL MOVEMENT ON
POPULATION DYNAMICS

Benjamin D. Dalziel, Ph.D.

Cornell University 2014

Many populations exhibit collective behavior, where interactions among nearby individuals scale up to cause emergent patterns in the behavior of groups, as in the coordinated movement of a flock of birds or a school of fish. Populations influenced by collective behavior violate the assumption of mass action that underlies most ecological models, in which individuals are viewed as statistically independent. However, the ecological significance of collective behavior is not well understood, because studies have been limited to populations where high throughput ethological data is available, such as in the laboratory or in computer simulations. This dissertation tests for the signal of collective behavior in ecological data—data on the distribution patterns of organisms collected on a coarser spatial and temporal scale than the underlying processes—and examines the influence of collective behavior on population dynamics. Data on the locations of migratory caribou (collected every five days by satellite tracking collars) are shown to be generated by two distinct processes. The first process creates broad-scale spatiotemporal order in movement patterns, and is likely driven by seasonally and spatially fluctuating environmental and physiological cues. The second process creates finer-scale order that is likely due to behavioral interactions among nearby individuals. The strength of alignment in the velocities of nearby individuals varies systematically with time of year, suggesting that collective behavior can be a dynamic property of migratory populations.

The dissertation then considers collective mobility patterns in humans, analyzing census data on the commuting patterns of workers in Canadian cities. The level of order in commuting patterns varies systematically among cities. In particular, in some cities a disproportionate number of workers travel to work in a few focal locations. Simulations of the spread of a respiratory infection in each city predict differences among cities in the risk of an epidemic, due to systematic variation in the level of order in the commuting patterns of workers. In particular, larger cities tend to be more highly organized and, as a result, have a disproportionately higher probability of sparking an epidemic. The dissertation then explores the role of large cities in supporting the emergence of a new strain of influenza in dogs. The analysis combines demographic data on animal shelters in the United States, molecular data from the pathogen and seroprevalence estimates from the literature to show that large animal shelters in major metropolitan areas function as endemic reservoirs for the virus, facilitating sporadic outbreaks in the wider population. In sum the dissertation research shows that collective behavior can sometimes be detected and characterized in ecological data without recourse to fine-scale behavioral observations, and that collective behavior can significantly alter population dynamics at broad spatial and temporal scales.

BIOGRAPHICAL SKETCH

Benjamin Dalziel was born in Ontario, Canada. He studied integrative biology at the University of Guelph, receiving a B.Sc. in 2004 and an M.Sc. in 2006.

ACKNOWLEDGEMENTS

My dissertation research was funded in part by an NSERC PGS-D grant from the Government of Canada, by a Canadian Institutes of Health Research grant (no. PTL-97126) to Babak Pourbohloul, by an Olin Fellowship from Cornell University, and by grants from the Andrew W. Mellon foundation. Thanks to the administrative staff in the Department of Ecology and Evolutionary Biology, and to staff in other offices at Cornell University, who provided logistical help. Thanks to the members of my dissertation committee—Stephen Ellner, Monica Geber, Giles Hooker and Colin Parrish—for the time and effort they devoted to improving the work. I am grateful to the people who collaborated on this work: Nimalan Arinaminpathy at Imperial College, Mael Le Corre and Steeve Côté at the University of Laval in Quebec, Edward Dubovi at Cornell University, Stephen Ellner at Cornell University, Bryan Grenfell at Princeton University, Jemma Geoghegan and Edward Holmes at the University of Sydney, Kai Huang and Colin Parrish at Cornell University, and Babak Pourbohloul at the University of British Columbia and the British Columbia Centre for Disease Control. Our collaborations form the “we” that narrates much of the dissertation, although any errors and omissions are my own. Stephen Ellner was the chair of my dissertation committee and a collaborator on all the research. His unfailingly lucid and constructive advice was always available to me. Above all I am grateful to my family, and especially to Jessica and Calvin Bliss. With them I have a wonderful life in which this work is a small but valued part, supported by their friendship and love.

TABLE OF CONTENTS

Biographical Sketch	iii
Acknowledgements	iv
Table of Contents	v
List of Tables	vii
List of Figures	viii
1 Introduction	1
2 Detecting collective decision making in ecological data on mobile animal groups	7
2.1 Introduction	7
2.2 Methods	12
2.2.1 Analytical approach	12
2.2.2 Collective behavior model	17
2.3 Results	21
2.4 Discussion	28
2.5 Acknowledgements	32
3 The dynamics of collective behavior in caribou migration	33
3.1 Introduction	33
3.2 Material and Methods	36
3.2.1 Analysis	37
3.2.2 Detecting the signature of collective behavior	40
3.3 Results	42
3.4 Discussion	50
3.5 Acknowledgements	54
4 Human mobility patterns predict divergent epidemic dynamics among cities	55
4.1 Introduction	55
4.2 Methods	57
4.3 Results	61
4.4 Discussion	64
4.5 Acknowledgements	67
5 Population dynamics, evolution, and control of emerging canine influenza virus in the United States	68
5.1 Introduction	68
5.2 Methods	72
5.2.1 Model	72
5.2.2 Vaccination - inactivated or modified live intranasal	75
5.2.3 Metapopulation dynamics	76
5.2.4 Phylogenetic analysis, estimates of R, and phylogeography	77

5.3	Results	80
5.3.1	Phylogenetic Structure of CIV in the USA	80
5.3.2	Epidemiological Dynamics of CIV in Shelter and Domestic Dogs	82
5.3.3	Populations that sustain viral transmission	87
5.3.4	Control and Eradication Strategies	89
5.4	Discussion	93
5.5	Acknowledgements	99
A	Supplementary Information for Chapter Three	100
A.1	Kernel-smoothing model	100
A.2	Properties of ψ for ensembles of random walkers	102
B	Supplementary Information for Chapter Four	104
B.1	Transportation models	104
B.2	Transmission model	105
C	Supplementary Information for Chapter Five	116
C.1	Mean field model of a single shelter	116
C.2	R_0 from mean field absent vaccination	117
C.3	Seroprevalence	120
C.4	Metapopulation model	120
	Bibliography	124

LIST OF TABLES

2.1	Simulation parameters	20
B.1	Summary of cities used in our study	110

LIST OF FIGURES

- 2.1 **Order in the movements of independent individuals exposed to environmental gradients of varying strength with noise level $\delta = 1$.** Horizontal axes in each pane comprise 100 points, with 100 replicates per point. Outer, lighter polygons extend vertically from the 5th to the 95th percentile; inner darker polygons encompass the interquartile range. **A:** Order parameter for populations of varying size in the absence of an environmental gradient ($\varepsilon = 0$). **B:** Order parameter for populations subject to varying strengths of environmental bias, for small (yellow) and large (purple) populations, sampled after 10 minutes (lower curve) and four hours (upper curve). Panes **C** and **D** show the same analysis as **A** and **B** using the adjusted order parameter. Variation in sampling time has no effect on the value of the adjusted order parameter for independent individuals, so the polygons for each of the two sampling periods are on top of one another. 16
- 2.2 **Contrasting collective intelligence with the aggregated responses of independent individuals using ecological data.** **A** and **B:** Distribution over space and time for model individuals released at the origin and heading upwards in a noisy environmental gradient. The cool colors (blues and greens) show 10 replicate model runs for nearly independent gradient followers ($\alpha = 32$). The hot colors show 10 replicate runs for populations that exhibit collective behavior ($\alpha = 0.5$). **C:** Distribution perpendicular to the gradient direction at the end of four hours for varying levels of collective behavior, with each replicate for a given value of α shown side-by-side. Boxes enclose the interquartile range and lines enclose the entire range of the data. Colors correspond to those in the previous panes. **D:** Distribution parallel to the gradient direction. 22

2.3	<p>Detecting collective behavior in ecological data. A: Adjusted order ($\bar{\psi}$) in the spatial distribution of simulated populations with different levels of collective behavior. Filled boxes enclose the interquartile ranges for the distributions of $\bar{\psi}$ across all sampling times and replicates. Lines extend from the 5th to 95th percentiles of the distributions. Hollow boxes show the analogous distribution for the null parameter $\bar{\psi}_0$. B: Adjusted order over time for populations strongly influenced by collective behavior (α from 0.125 to 0.5; triangles) and weakly influenced by collective behavior (α from 8 to 32; circles), averaged over replicates. Colors correspond to those A. C: Average adjusted order over time for populations with intermediate levels of collective behavior ($\alpha = 1$, dotted line and crosses; $\alpha = 2$, solid line and squares; $\alpha = 4$, solid line and diamonds.</p>	25
2.4	<p>Behavior of the adjusted order parameter when some individuals are unobserved. Lines show the average value of $\bar{\psi}$ across replicates for populations showing collective behavior ($\alpha = 0.5$; warm colors) and populations of nearly independent gradient followers ($\alpha = 32$; cool colors), when a proportion of the individuals are randomly removed from the analysis. Each of the 10 lines of a given color and style corresponds to a repetition of the analysis on a different replicate simulation. Each line is composed of 30 points, which are each derived from a different random sample of the original simulation output.</p>	27
2.5	<p>Detecting collective decision making in a temporally shifting environmental gradient. A: Direction of travel over time for a population with $\alpha = 0.5$. Vertical lines encompass the range of directions of travel for individuals observed at a particular time. The thick grey curves shows the true direction of the gradient. The red line shows a cubic spline fit to the velocity data. B: Adjusted order and null order over time in a second replicate realization of the process, where velocities were adjusted by subtracting the spline fit to the velocity data in the first replicate (as shown in A). The inset in B shows the analogous results for a population of nearly independent gradient followers ($\alpha = 32$). C shows the distribution of the horizontal component of the adjusted velocities in the second experiment, showing collective deviations from the true gradient direction, as estimated in the first experiment.</p>	31

3.1	<p>Migration patterns of the Rivière-aux-Feuilles caribou herd. The inset globe shows location of the study area. The two larger maps show the locations and velocities of GPS-collared caribou observed during the spring (May) and and fall (October) migrations, pooled across years from 2003 to 2011. The style of the points show each caribou’s velocity, according to the legend on the left. Bottom panels show time series of individual locations and velocity over the study period. Dark regions enclose the interquartile range; lighter regions enclose the 5th to 95th percentiles, for running quantiles with non-overlapping adjacent windows. The window widths were 30 days for location and seven days for velocity.</p>	44
3.2	<p>Performance of the advection-diffusion model of caribou migration, simulated by releasing independent particles into its velocity field. Panes show easting (a) and northing (b) as a function of time of year, for caribou (points and lines) and particles (shaded area). Points show running median caribou locations using seven day consecutive non-overlapping windows. Vertical lines enclosed the interquartile range for the same windows. The shaded region encloses the analogous interquartile range for the particles.</p>	47
3.3	<p>The signature of collective behavior in caribou migration patterns. Top panel: time series, modulo year, of the order parameter ψ for the spatio-temporal neighborhoods associated with collective behavior (blue, solid line) and advection-diffusion (red, dashed line). Lines show the medians, darker regions enclose the interquartile range, and lighter regions enclose the 5th to 95th percentiles, for running quantiles with 7 day non-overlapping adjacent windows. Bottom panel: the same analysis done on simulations where caribou do not interact.</p>	48
3.4	<p>A spike in collective behavior of migrating caribou each year in July. (a) Points along the horizontal axis show the average date of calving (b) Latitude of the the caribou for the same time range; the spike in collective behavior seems to be right before they arrive at their farthest north location, and when herd density is high.</p>	49

- 4.1 **Mobility patterns of workers in cities.** The thickness and color of edges show the number of individuals commuting between census tracts (CTs). Circles are actually short edges, representing individuals who live and work in the same CT. Larger cities tend to have more highly organized commuting patterns, as measured by the average number of workers who have their workstation in the same CT as a randomly chosen worker (m^*). However, cities also show marked differences in organization that are independent of population size 60
- 4.2 **Systematic differences in worker mobility patterns among 48 Canadian cities. A:** \bar{m} , the mean number of workers per census tract (CT) (triangles) and m^* , the average number of workers in the same CT as a randomly chosen worker (circles), as a function of population size (N). The solid line shows \hat{m} , the fitted relationship between m^* and N . The vertical distance between the dashed lines spans $\frac{2\sigma}{\sqrt{\pi}}$, where σ is the standard deviation of m^*/\hat{m} , showing the expected absolute difference in m^* (on a \log_{10} scale) between two cities of the same size. The width of the shaded polygon then shows what change in N would produce that difference according to \hat{m} . **B:** Variance explained in each city by the configuration (squares) and radiation (diamonds) models of commuting flows. 63
- 4.3 **Epidemic dynamics as a function of heterogeneity in human mobility patterns. A:** Probability that a single infection will spark an epidemic in 48 cities with different levels of organization in their commuting patterns, calculated from 100 simulations for each city for transmissibilities of $\lambda = 1$ (triangles) and $\lambda = 10$ (circles). Point size is proportional to $\log_{10} N$. Lines show logistic regression controlling for transmissibility. **B:** Relative risk of an epidemic as a function of excess heterogeneity in mobility patterns. The statistical model for \hat{P} is $\text{logit}\hat{P} = x_\lambda \log N + w_\lambda$. Lines show linear regression controlling for transmissibility. **C:** Final number infected is positively correlated with the level of heterogeneity in mobility patterns. Lines show fits of linear regression on log-transformed variables; $\lambda = 1$ (triangles, dashed line), $\lambda = 10$ (circles, solid line). Point size is proportional to $\log_{10} N$. **D:** This effect persists when the effects of population size on F and m^* are removed. 65

5.1	Phylogenetic trees of HA1, NP and M sequences for EIV (black) and CIV (colors).	Boxes surround CIV clades comprising two or more samples from the same US state. Branches leading to CIV samples from the same location are colored by location (New York, blue; Pennsylvania, orange; Colorado, red). Branches leading to CIV samples from multiple locations are colored grey.	81
5.2	Demography of dogs in US animal shelters. A:	Cumulative distribution of median population size in each shelter (dashed line) compared to a negative binomial distribution fitted to the data (solid line). B: Intake rate as a function of population size. Points show the median value for each shelter and vertical lines enclose the interquartile range. Line shows fit by linear regression to log-transformed median intake rates. C Length of stay as a function of shelter size. The slope of the dashed line does not differ significantly from 0. D Cumulative distribution of length of stay across all shelters (bars) compared to an exponential distribution with mean rate $1/9.88$ days ⁻¹ (solid line).	83
5.3	Seroprevalence, R_0 and R for CIV, estimated from host demographic data, seroprevalence data, and molecular data. A:	Saturating relationship between seroprevalence and R_0 in a stochastic SIR framework, parameterized from the shelter intake and output data. Red line shows equilibrium seroprevalence predicted by the mean-field model. Points show point seroprevalence estimates from the stochastic simulations, where 74 dogs are sampled at random in a shelter with an average dog population of 134, corresponding to [152]. B: Deviations of point seroprevalence estimates from the long-term average (bars) compared to a normal distribution (line). C: Posterior distribution of R_0 based on an observed seroprevalence of 0.41 in [152]. D: R for CIV, estimated by fitting a birth-death skyline phylodynamic model to HA1 gene sequences. The black line shows the mean estimate while the grey shaded shows the highest probability density (HPD) range, encompassing 95% of the credible set of sampled values.	86

5.4	Demographics, persistence, spread rate and possible eradication of CIV.	A: Dog population sizes in animal shelters and within-shelter spread rates at which CIV can persist for at least 100 days according to present intake and output rates. The surface shows a smoothed version of the outcome of 1000 simulations conducted at random points within the plane described by the figure. Darker shades correspond to higher probabilities of persistence. Red symbols show features of the empirical joint distribution for dog population size and R_0 in shelters (see Figures 5.1 and 5.2), including the median (hollow circle), mean (filled circle), 2.5th percentile (minus sign) and 97.5th percentile (plus sign). B: Results of an intervention that reduces the arrival rate of susceptible individuals at a shelter to 1/3.9 its current value, equivalent to reducing the mean estimate for R_0 to 1.	88
5.5	Predicted performance of a control program using a live-attenuated vaccine administered to dogs on arrival in US animal shelters.	A: A vaccine that removes individuals from the chain of transmission with 85% probability ($\kappa=0.15$) within 24h ($\alpha=1$ day) is predicted to eradicate CIV from shelters within six months. The simulations used 100 shelters with dog population size, intake rate, and outtake rate jointly sampled with replacement from the shelter demographics data, and $R_0 = 3.9$. White lines show medians and shaded areas enclose the 5th to the 95th percentiles of the simulation data. B: Decreasing vaccine efficacy to 75% can still achieve eradication in isolated shelters (blue region, solid line), however shelters that transfer dogs amongst themselves at the observed mean rate of $\tau = 0.1$ would preserve CIV in a few shelters despite the vaccination program (red region, dashed line). C: Further decreases in vaccine efficacy make eradication significantly less likely, particularly if shelters are connected through the transfer of dogs.	90

5.6	<p>A CIV invasion over multiple shelters, starting with an infection in a single large shelter. A: Each vertex represents an animal shelter with dog population size proportional to the area of the circle. Edges show transfer of infection from shelter to shelter over time through the movement of infected dogs. Edge lengths are arbitrary. The data for this figure were produced by simulating the metapopulation stochastic <i>SIR</i> model with 100 shelters for 100 days, starting with a single infection in the largest shelter. Population sizes were sampled with replacement from the shelter data. $R_0 = 3.9$. Transfer probability is set to the mean observed value of $\tau = 0.1$. B: Large shelters tend to receive the infection earlier (and more often) following an outbreak at another shelter. C: Probability that CIV will persist for 100 days in a shelter of a given size following the introduction of a single infected individual to an otherwise susceptible population. . . .</p>	92
B.1	<p>Fit of the gravity model in each city as a function of population size. We used alternate distance weighting functions ($f(r)$) in the denominator. a: shows $f(r) = r^\delta$. b: shows $f(r) = \exp(\delta r)$</p>	113
B.2	<p>Comparing the values of epidemic probes from the simulations—the probability that a single initial infection will spark an epidemic, peak attack rate and final attack rate—across different network topologies—real, no home-work correlations, and configuration model—for $\lambda = 1$ (triangles) and $\lambda = 10$ (circles).</p>	114
B.3	<p>R_0 on the configuration model network as a function of λ. Points show median values across cities and simulation runs; vertical lines go from the first quartile to the third quartile of the distribution across cities and simulation runs.</p>	115

CHAPTER 1

INTRODUCTION

A population of organisms is defined by the potential for its constituents to interact. These interactions play an important role in ecological and evolutionary dynamics, encompassing sexual reproduction, resource competition, and the transmission of infectious diseases. Many standard models of population dynamics, including those of Malthus [1], Verhulst [2], Lotka and Volterra [3, 4], Kermack and McKendrick [5], Holling [6], and Rosenzweig and MacArthur [7], correspond to the general form

$$\frac{dx}{dt} = xg(x, y) \tag{1.1}$$

where the influence of the individual behaviors that determine population dynamics are simply described by a per-capita population growth rate function g . This per-capita rate is determined by current population size $x(t)$ and other time-varying quantities, $y(t)$, including the sizes of other populations, or environmental conditions such as temperature or nutrient availability.

Despite their prevalence and apparent generality, these models make a strong assumption about how interactions among individuals affect the dynamics of populations. The assumption is that it does not matter which individuals interact. Rather population growth is assumed to depend only on the average rate of interaction, which is determined solely by x and y . This is the mean field approach to modeling population dynamics, originating from the law of mass action first described by Waage and Gulberg in 1864 for chemical reactions [8, 9].

By focusing on average interaction rates instead of individual interactions, mean field models act as if individual behavior is independent and identically

distributed. This greatly simplifies theory [9], and has been productive to the extent that some mean field models have been suggested to represent ecological laws [10]. The mean field approach has been successfully applied in many scenarios, including spatial dynamics [11] such as the spread of advantageous mutations [12] or invading organisms [13], the dynamics of age or stage structured populations [14, 15], and interactions between ecological and evolutionary dynamics such as in the evolution of trait-mediated defense in predator-prey systems [16] and in the spread and evolution of infectious pathogens [17].

Mean field models may succeed for three reasons. First, individuals may be independent in a statistical sense despite their interactions with one another. Interacting individuals are not guaranteed to achieve a correlation in behavior sufficient to separate a population's trajectory from the predications of a mean field model. Second, populations of independent individuals can still generate cohesive patterns. For example, diffusion is characterized by a cohesive pattern among large numbers of particles, originating from the independent brownian motion of each one [18]. So mean field models can make a range of predictions that can be tested with data from real populations. Finally, mean field theory does not simply ignore the complexity of individual interactions, rather it attempts to find simpler models that accurately describe the net outcome of these interactions. For example, mean field models can include more state variables to capture higher statistical moments of the spatial distribution of a population (such as variance in density over space). In some cases, these higher moments can be described as deterministic functions of the lower ones, allowing an exact moment closure that produces a simple and reliable mean field model [19]. More generally, reliable simpler models for classic high-dimensional systems do exist, such as the diffusion equation for the stochastic movement of independent

particles [20].

However, there are conditions that are common in biological populations under which a mean field approach yields misleading results. For example, infectious contact networks—expressing which pairs of individuals have contact that would be capable of transmitting a disease—built by simple algorithms for individual behavior, can have disease dynamics that depend on their topology, and not only on the mean per-capita transmission rate [21, 22, 23, 24]. Another example concerns evolutionary dynamics in structured populations where individuals are selectively replaced with their neighbors' offspring [25]. Rules for who can replace whom are given by a directed graph, where individuals are vertices and neighbor relationships are edges. When a mutant appears in the population, the topology of the graph can amplify either selection (increasing the probability that an advantageous mutation will become fixed in the population), or drift (favoring the fixation of random mutations). Crucially, selection or drift amplification occur more readily in populations whose graph is more highly organized (in the sense of a statistically significant departure from a homogeneous random graph) [25]. More generally, self organization is a defining feature of life [26]. And so it is in some ways surprising that we still use essentially the same approach to predict a simple chemical reaction in a flask as we use to explain predator-prey dynamics of moose and wolves [27], or the spread of measles within and among cities [28].

One type of self organization in populations that has received increasing attention is collective behavior, where the activities of each individual is influenced by others nearby, causing the population to adopt a more cohesive pattern than would be expected for a group of independent individuals [29]. Typically

the activity involved is movement, so that nearby individuals align their velocities, but the process is generalizable to other changes in state besides location. Collective behavior can cause a population to exhibit behaviors that would be impossible for ensembles of independent individuals to achieve. Animal groups can make collective movement decisions according to information held by only a few individuals [30]. Slime molds (*Physarum polycephalum*), which have no individual capacity to learn, can collectively predict future environmental disturbance from past experience [31]. And autonomous internet users can collectively estimate the burden of influenza-like illness based on personal interest in flu symptoms [32].

Collective behavior can arise from simple rules for individual behavior, suggesting it could be widespread in nature [33, 34]. Yet to date few studies have addressed how to detect collective behavior outside of the laboratory, or explored how populations influenced by collective behavior might differ in their ecological and evolutionary dynamics from the predictions of mean field models. This dissertation is interested in detecting collective behavior in ecological data (data on the distribution and abundance patterns of organisms that is collected on a coarser scale than the underlying processes) and examining its potential consequences for population dynamics.

Chapter two addresses the inverse problem of distinguishing collective behavior from the aggregated responses of independent individuals, when the causal behaviors are not observed [35]. We use a well-studied model of collective movement [36, 30] sampled at a coarser scale than the underlying behavior processes to show how collective deviations from the average direction of travel have a high positive predictive value for collective movement.

Chapter three examines the role of collective behavior in data on caribou migration patterns, using measurements of the location of the Leaf River caribou herd in Québec, Canada, collected using satellite tracking collars. Are migration patterns driven primarily by physiological and environmental cues available to each individual directly? Or is migration in part an emergent property of social interactions among individuals, particularly the tendency for nearby individuals to align their velocities? We find that a model of migration where independent individuals choose their velocities based on location and time of year can reproduce the observed movement patterns of the herd. However, residual variation in velocities that cannot be explained by the model show a hallmark of collective behavior - high polarization order in the velocities of nearby individuals. Comparing the amount of polarization observed to that predicted by the model at various locations and times of year suggests predictable seasonal fluctuations in the strength of collective behavior. This implies collective behavior influences caribou movement patterns in a way that changes systematically over space and time.

Chapter four turns to humans, where cities represent a classic example of self-organization that may result in part from collective behavior [26]. Here we explore one aspect of that organization, using census data on the mobility patterns of workers in Canadian cities. A fundamental prediction of contact network epidemiology (a branch of ecology and epidemiology that is outside the mean-field approach) is that heterogeneity in individual behavior can lead to systematic differences in disease dynamics among host populations. We use the census data to estimate contact networks for 48 Canadian cities, finding both size-dependent and size-independent systematic differences in mobility patterns. These are predicted to lead to significant disparities among cities in

the dynamics of respiratory infections. In particular, larger cities are predicted to have significantly higher risks of sparking an epidemic, due to their more highly organized commuting patterns.

Chapter five examines the role of cities in affecting the emergence of a novel strain of influenza, analyzing ecological and evolutionary data on the epidemiology and phylogeny of canine influenza virus (CIV). We find that CIV is maintained in a few large animal shelters associated with major metropolitan areas. These shelters function as endemic reservoirs for CIV that serve as staging grounds for sporadic outbreaks in the wider population, and rescue the virus from extinction due to demographic stochasticity. The role of metropolitan animal shelters in the epidemic dynamics of CIV is analogous to the role predicted for specialized work areas in the mobility patterns of workers in larger cities, because these structures efficiently bring large numbers of susceptible and infected hosts into close proximity. The results of the CIV study are thus consistent with the predictions of the analysis of commuting patterns in cities, supporting the hypothesis that large cities can disproportionately influence disease dynamics by coordinating the movements of hosts.

CHAPTER 2
DETECTING COLLECTIVE DECISION MAKING IN ECOLOGICAL
DATA ON MOBILE ANIMAL GROUPS

2.1 Introduction

Many populations exhibit collective behavior, where localized interactions among neighboring individuals lead to broad scale patterns in the behavior of groups, as in the coordinated movement of a flock of birds or a school of fish [29, 37]. Populations influenced by collective behavior violate the assumption of mass action that underlies most ecological models, in which individuals are viewed as statistically independent [9, 19]. Correspondingly, collective behavior can allow groups to track variable resources more effectively than independent individuals [38, 39, 40, 41, 34], leading to increased fitness through population-level cognitive responses to variable environments [42, 43, 44, 45].

Research on collective decision making has advanced by identifying the underlying behavioral processes that govern interactions among neighboring individuals in a population [46, 47, 48, 35, 34, 49], with data-driven analyses using fine scale ethological observations to infer how strongly, and in what ways, individuals are influenced by social interactions [50, 47, 48]. The behavioral rules revealed by these analyses are often simple, requiring only rudimentary cognitive abilities [34]. This suggests that collective behavior could be a widespread adaptation to life in an uncertain world.

Yet few populations are intensively sampled at the scale of individual behavioral decisions. As a result, collective behavior may be more widespread

in nature than our current ability to detect it. In this paper we suggest an approach to screen for collective behavior in ecological data—data on the distribution patterns of organisms collected at a coarser scale than the underlying behavioral processes [51]—to help identify systems where more detailed studies of the role of collective behavior may be fruitful. As the information we have on most populations is sparsely sampled—excluding most individuals, in most places, at most times—the approach we propose focuses on detecting features of collective behavior that are robust to changes in the details of the underlying individual interactions [52, 48], addressing a simple case of the inverse problem of distinguishing collective behavior from the aggregated responses of independent individuals, when the causal behaviors are not observed [35].

Simulations show that simple rules can lead to the emergence of collective behavior. Individuals that aggregate can pool their estimates of a noisy environmental gradient allowing improved navigation for the group [39]. Aligning velocities with neighbors can also improve the navigational abilities of simulated groups [38, 40, 41]. In this case only some individuals are required to have information about the environment in order to facilitate a collective response, and the informed leaders do not need to be distinguishable by the others [30, 53].

Laboratory and field studies reveal how mechanisms of collective behavior identified *in silico* play out in real populations. For example, golden shiners (*Notemigonus crysoleucas*) swimming in a tank with spatially varying light levels reduce their speed in darker areas, as well as heading toward other individuals who are nearby [34]. As a result the population remains in the darker areas of the environment more effectively than ensembles of independent individuals [34]. Predator-prey interactions observed in fishes show how collective tactics

of capture and escape are employed, and disrupted, by both predator and prey populations [43, 44].

Along with fishes, social insects provide many of the current examples of collective behavior [54]. But there is evidence in other systems too, ranging from slime molds (*Physarum polycephalum*) that collectively predict environmental fluctuations from past experience [31] to autonomous internet users who collectively perform disease surveillance [32]. These populations achieve a type of cognition that would be impossible for a single individual. In the case of the slime molds, this involves using collective memory of past environmental disturbances to predict future ones, even though no single individual can remember the environment in this way [31]. In the case of the internet users, collective intelligence emerges when aggregated data on personal interest in influenza symptoms accurately estimates the incidence of influenza-like illness [32].

And yet despite the potential importance of distributed emergent cognition, populations often respond to information in their environment that individuals can detect and respond to individually. For example, fish schools that collectively evade predator attacks are composed of individuals who can independently detect and respond to the presence of a predator, although the population's collective responses may be more rapid and effective than those of independent individuals [44]. In cases where independent individuals have the capacity to respond directly to the stimuli in question, there is, then, a basic challenge of distinguishing collective behavior from the aggregated responses of independent individuals [55]. This is made more difficult if the underlying behaviors are not observed.

Our approach to this challenge is based on testing for between-individual

correlations that would be unlikely if individuals were making completely independent responses to information in their common environment. To motivate the approach, consider the problems of detecting plagiarism or cheating on a written test, without observing the subjects' behavior. As with ecological data, we often have information on outcomes (such as test results, or a submitted paper) rather than observations of the underlying behaviors. Plagiarism is often detected from similarity to an originating work that would be unlikely in the absence of direct copying. Cheating can be detected from suspiciously high similarity between two individuals' answers, such as a statistically improbable sequence of identical right and wrong answers on a multiple choice test. Regardless of how they cheated, or what the test was about, improbably similar outcomes imply a low likelihood of independence.

As another example, consider detecting collective behavior in data on the velocities of cars on a section of highway, collected over replicate days. A controlled experiment where we repeatedly and randomly select the trajectory data from two cars—either cars from the same day, or cars from different days—and record whether or not the trajectories of the cars collide, would reveal that the risk of collision for cars selected from different days is higher than for cars from the same day. Without detailed observations of the behaviors of the cars, this experiment allows us to exclude the possibility that the data were generated by replicate realizations of a process on independent individuals. If individuals ignored each other, the trajectories of cars from the same day would collide just as often as the trajectories of cars from different days.

A recently described approach for detecting the influence of conspecifics on population distribution patterns uses a similar logic to our second example,

testing how individual displacement patterns differ from an explicit null model where individuals are independent, by testing the extent to which the proximity between individuals can be attributed to independent random displacements [56]. Like [56], our approach uses a comparison with what independent individuals would do as a basis for detecting collective behavior. However, unlike [56], our approach does not require specifying a null model for independent individuals, nor does it seek to parameterize a particular statistical model for analyzing collective behavior. Instead we focus on emergent patterns that are fundamental to collective behavior, inherently unlikely for ensembles of independent individuals, and robust to unobserved individuals, as well as to data observed at much coarser scale than their underlying movement decisions. Our approach focuses on animal movement data, but it generalizes straightforwardly to other movement data and to other kinds of changes in individual state.

We use a well-studied model of collective animal movement [30] implemented in an environment that has a gradient which represents the population's preferred direction of travel. But the environment is noisy, so that at each place and time an individual's experience of the gradient varies. Varying the strength of social interaction among nearby individuals reveals that the appearance of broad scale order is not diagnostic of collective behavior, because independent gradient followers also tend to be highly ordered, even in noisy environments. However, populations influenced by collective behavior show broad scale collective idiosyncratic deviations from the true gradient direction that are visible in sparsely sampled data, and are statistically unlikely for independent individuals, regardless of the behavioral rules that independently govern each of their trajectories. We argue that these "collective mistakes" represent a characteristic feature of collective behavior in ecological data.

2.2 Methods

2.2.1 Analytical approach

The approach we propose considers a population of N individuals whose locations change continuously in space and time. However individual locations are only observed at M discrete time points t_1, t_2, \dots, t_M . Sampling periods $s_j = t_j - t_{j-1}$ measure the temporal separation between “bouts” of observation, wherein each individual’s location is recorded. Let $x_i(t_j)$ represent the location of the i th individual observed at time t_j . The velocity of an individual associated with a pair of adjacent sampling times is estimated as

$$v_i(t_j, s_j) = \frac{x_i(t_j) - x_i(t_{j-1})}{s_j}, j > 1 \quad (2.1)$$

and thus can vary with different sampling periods, as well as over time and among individuals.

A well-studied measure of collective movement is the order parameter

$$\psi(t_j, s_j) = \frac{1}{N} \left| \sum_{i=1}^N \frac{v_i(t_j, s_j)}{|v_i(t_j, s_j)|} \right| \quad (2.2)$$

which represents the average normalized velocity of the population at time t_j . In the limit as N becomes large $\psi(t_j, s_j)$ ranges from 0, when individual velocities have uniform random directions, to 1, when individual velocities all have the same direction [37].

The sampling period s is typically < 1 second in ethological studies of collective behavior [30, 46, 34]. However in ecological data s will typically be much larger, as the spatial distribution of organisms usually is sampled at discrete and relatively distant times in ecological studies, rather than in nearly continuous

time. Correspondingly, we now consider the behavior of the order parameter for finite populations whose positions are sampled at discrete time points that can be arbitrarily distant.

We begin with the null case where individuals velocities are independent of one another. Suppose individuals move in a two-dimensional environment that has a gradient with direction vector $\phi = (0, 1)$ which represents the preferred direction of travel for the population. The changing spatial distribution of the population over time then follows an advection-diffusion model

$$\frac{\partial n}{\partial t} = \delta \nabla^2 n - \varepsilon \phi \cdot \nabla n \quad (2.3)$$

where $n(x, t)$ is the density of the population at location x and time t , and ∇ represents the gradient of $n(x, t)$ [13, 57, 11, 20].

The advection parameter ε represents the strength of the environmental gradient. As ε increases individual velocities become increasingly aligned in the preferred direction of travel, all else equal. The parameter δ controls the rate of diffusion, representing random movement not associated with following the preferred direction. Higher values of δ can represent a “noisier” environment, where each individual’s estimate of the preferred direction of travel at a given time is increasingly uncertain. If a population of independent individuals is released at the point $(0, 0)$ and observed s seconds later, their spatial distribution will follow a bivariate Gaussian distribution with mean $\varepsilon \phi s = (0, \varepsilon s)$ and variance-covariance matrix $s\delta I$, where I is the identity matrix. Note that variance in location grows linearly with time in advection diffusion, which is why the variance-covariance matrix has a factor of s .

We calculated the order parameter $\psi(t_j, s_j)$ on simulated data from Equation 2.3 for a range of population sizes, strengths of environmental bias, and sam-

pling periods. This demonstrates the intuitive result that ecological data on the mobility patterns of independent individuals can display highly ordered velocities without any collective behavior. This higher order can be generated by chance in smaller populations, where there is a higher probability that random velocities will be aligned (Figure 2.1A), by a strong environmental gradient that dominates the effect of noise, and by longer sampling intervals which reduce the effect of noise by averaging it over a long time period (Figure 2.1B). Detecting collective behavior in ecological data therefore requires a different statistic, one that attains distinctly different values for groups of independent individuals than for collectives.

Consider an adjusted velocity

$$\bar{v}_i(t_j, s_j) = v_i(t_j, s_j) - v_0(t_j, s_j) \quad (2.4)$$

where v_0 represents an expectation for v_i if individuals were acting independently. As a model for what independent individuals would do, v_0 could be complex and we return to the issue of determining v_0 below, addressing in particular the case when the preferred direction of travel, ϕ , varies over time. To demonstrate our approach we focus on a simpler case where ϕ does not change during the time the population is observed. In that case the expected velocity of an independent individual at a certain time, given a set of observed velocities (which may or may not be from independent individuals) is just the overall average observed velocity

$$v_0 = \frac{1}{MN} \sum_{j=1}^M \sum_{i=1}^N v_i(t_j, s_j) \quad (2.5)$$

across all time points. In other words, if ϕ is constant, then the overall mean velocity over time is a good model for the velocity of any particular independent individual, at any time. This works because independent individuals do not

interact with one another and so their velocities are exchangeable across time points.

We now introduce an adjusted order parameter

$$\bar{\psi} = \psi|_{v_i \rightarrow \bar{v}_i} \quad (2.6)$$

which is analogous to the one in Equation 2.2, except that it is calculated on the adjusted velocities \bar{v}_i instead of the raw velocities v_i . As with velocities v_i and the order parameter ψ , the adjusted order parameter $\bar{\psi}$ is a function of time t and sampling period s . However we will sometimes write these functions with their arguments suppressed for visual clarity. When we do so, comparisons of the value of a function under different circumstances (e.g. "the value of $\bar{\psi}$ is higher than... ") imply that the comparison is being done at an arbitrary time point and sampling period, unless otherwise indicated (Figure 2.2C).

The value of the $\bar{\psi}$ tends to be lower than ψ for independent individuals, meaning that adjusted velocities, \bar{v}_i , are less ordered than raw velocities v_i for independent individuals. This is because the subtraction of v_0 in Equation 2.4 removes some of the order that is due to individuals heading in the same direction independently—either by random chance, or because of exposure to a common environmental gradient. As a result, $\bar{\psi}$ is less influenced by population size (Figure 2.1C), and is unaffected by the strength of the environmental gradient or the length of the sampling interval s (Figure 2.1D).

Whereas $\bar{\psi}$ tends to be low for ecological data on independent individuals, regardless of the population size, the strength of environmental bias, or the sampling period, we hypothesize that increasing levels of collective behavior lead to increases in the value of $\bar{\psi}$. Increased values of $\bar{\psi}$ for populations influenced by collective behavior are not due to increases in their ability to travel in the true

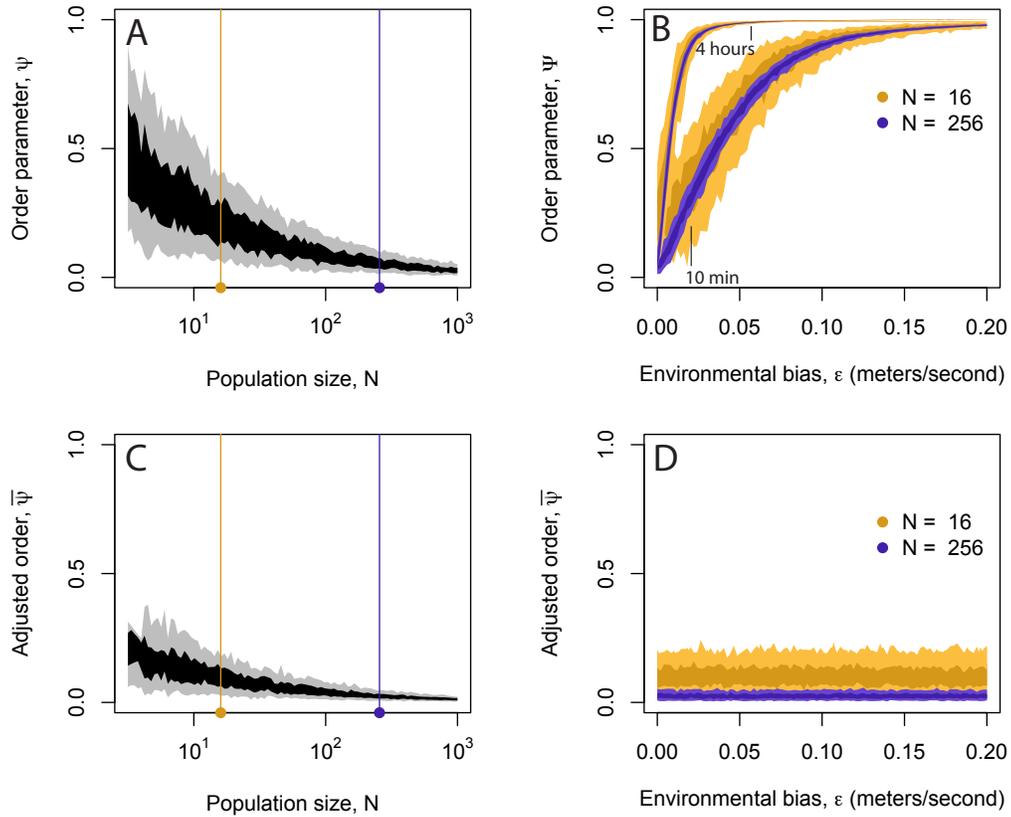


Figure 2.1: **Order in the movements of independent individuals exposed to environmental gradients of varying strength with noise level $\delta = 1$.** Horizontal axes in each pane comprise 100 points, with 100 replicates per point. Outer, lighter polygons extend vertically from the 5th to the 95th percentile; inner darker polygons encompass the interquartile range. **A:** Order parameter for populations of varying size in the absence of an environmental gradient ($\varepsilon = 0$). **B:** Order parameter for populations subject to varying strengths of environmental bias, for small (yellow) and large (purple) populations, sampled after 10 minutes (lower curve) and four hours (upper curve). Panes **C** and **D** show the same analysis as **A** and **B** using the adjusted order parameter. Variation in sampling time has no effect on the value of the adjusted order parameter for independent individuals, so the polygons for each of the two sampling periods are on top of one another.

gradient direction (the effect of traveling further up the gradient is removed by the subtraction of v_0 in Equation 2.4), or because collective behavior increases alignment in the gradient direction (because an improved overall alignment is also captured by v_0). Rather, the proposed mechanism is that at any moment, a population influenced by collective behavior will have its own idiosyncratic deviation from motion in the true gradient direction, due to the propagation to larger scales of stochastic interactions between nearby individuals. We will show that these “collective mistakes” can allow us to detect collective behavior in sparsely sampled ecological data.

2.2.2 Collective behavior model

To demonstrate our method for detecting group decision making in ecological data, we simulated data from well-studied model for collective behavior [30]. The model is implemented in an unbounded environment and observed over a long period relative to the time step of the simulation (the model steps forward 0.2 s at a time and we observe it for 4 simulation hours). As above, the environment has a gradient with a constant mean direction, and uniform noise that is independent and identically distributed over space and time. As above, the environmental gradient represents the preferred direction of travel for individuals.

At each time step, individual velocities in the model are given by

$$v_i(t+h) = \langle v(t) \rangle_i + \alpha g_i + z_i \quad (2.7)$$

followed by rescaling to unit length

$$v_i(t+h) \rightarrow \frac{v_i(t+h)}{|v_i(t+h)|} \quad (2.8)$$

where the vector $v_i(t)$ is the velocity of the i th individual at time t , and the time step of the model is h . $\langle v(t) \rangle_i$ represents the velocity chosen by i in response to the positions and velocities of its neighbors, as detailed below. Individuals are constrained by a maximum turning angle, such that the interior angle between $v_i(t+h)$ and $v_i(t)$ can be at most θ_{max} .

The random variable g_i represents the preferred direction of travel as it is perceived by individual i at time t . As above we assume a two-dimensional world in which the true preferred direction is the vector ϕ . Each time step an individual has access to a noisy estimate of the gradient that has unit magnitude and deviates from the true direction by an angle θ_g , which is uniformly distributed on the interval $(-\sigma_g, \sigma_g)$. An individual weighs g_i in their final desired velocity according to the gradient response parameter α . When α becomes large, individuals move independently from one another.

The vector z_i represents random error in velocity. z_i is a randomly chosen point on a circle centered at $(0,0)$ with radius σ_z . The larger the value of σ_z , the less an individual's velocity is based on cognitive responses to the environment or to the locations and velocities of its neighbors. As σ_z becomes large, each individual performs a random walk.

$\langle v(t) \rangle_i$ is chosen based on the locations and velocities of i 's neighbors as follows. Each time step, an individual's first priority is collision avoidance. If there are other individuals within the ball representing the focal individual's zone of avoidance, with radius r_a , then $\langle v(t) \rangle_i$ points away from the mean direction to those individuals.

If there are no individuals within the focal individual's zone of avoidance, then $\langle v(t) \rangle_i$ is based on the positions and velocities of neighbors within the zone of social interaction, a ball with radius $r_s > r_a$. $\langle v(t) \rangle_i$ is then the average of the vector toward the centroid of i 's neighbors, and the vector representing the mean velocity of those neighbors. $\langle v(t) \rangle_i$ is always normalized to have unit magnitude.

An individual's position changes over time according to

$$x_i(t+h) = x_i(t) + f v_i(t+h) \quad (2.9)$$

where f is the speed of each individual. Note that spatial variation is implicit in the model because at each time t , individual i is at a specific location $x_i(t)$. The preferred direction in the environment, and individuals' perceptions of their neighbors' locations and velocities thus vary spatially as well as temporally.

To summarize the model, simulated populations attempt to follow a noisy environmental gradient with constant mean direction using a mixture of individual- and group-level cognitive responses. The balance between the two types of cognition is determined by the gradient response parameter α , with increasing values of α representing increasing independence among individuals. The parameterizations we used are shown in Table 1, and follow [30]. In each simulation we obtained broader scale ecological data by recording the spatial distribution of the population every 10 minutes for 4 hours.

Table 2.1: Simulation parameters

Parameter	Interpretation	Value
h	Time step	0.2 seconds
r_a	Radius of avoidance	1 meter
r_s	Radius of social interaction	6 meters
f	Speed	1 meter / second
θ_{max}	Maximum turning angle	2 rad
θ_g	Environmental noise	2.5
θ_z	Individual noise	0.02
N	Population size	256
M	Number of replicates	10
s	Sampling period	10 minutes
α	Gradient response parameter	0.125, 0.25, 0.5, 1, 2, 4, 8, 16, 32
Initial positions	Uniform within a 30m x 30m square	

2.3 Results

At the maximum value of the gradient response parameter we examined ($\alpha = 32$) individual velocities are nearly independent. Correspondingly, populations at that level of α follow the spatiotemporal patterns predicted by advection diffusion (Figure 2.2A,C,D). In particular, in multiple replicate runs of the model, populations of independent individuals tend to the same broad-scale spatial distribution in all replicates, because the velocities of independent individuals are exchangeable over space and time. In contrast, lower values of the gradient response parameter lead to systematic differences in velocity over time and among replicates (Figure 2.2B). These systematic differences are driven by social interactions among neighboring individuals that scale up to cause population-level idiosyncratic deviations from the preferred direction of travel—“collective mistakes” (Figure 2.2B,C). At the same time, scaling up local conspecific interactions is what advantages collectives over independent individuals in variable environments, enabling populations with collective behavior to travel more quickly and precisely in the preferred direction of travel (Figure 2.2D). This effect persists until values of α become so low that individuals cease to respond much to the gradient. In this case the population still “drifts” in the direction of the gradient, while maintaining a highly heterogeneous spatial distribution (e.g. $\alpha = 0.125$ in Figure 2.2C,D).

Because populations of independent individuals also align to follow the gradient, alignment of movements in the gradient direction is not sufficient evidence for collective behavior in ecological data. However, alignments that involve broad-scale group-level deviations from the mean gradient direction do have a higher predictive value for identifying collective behavior, particularly

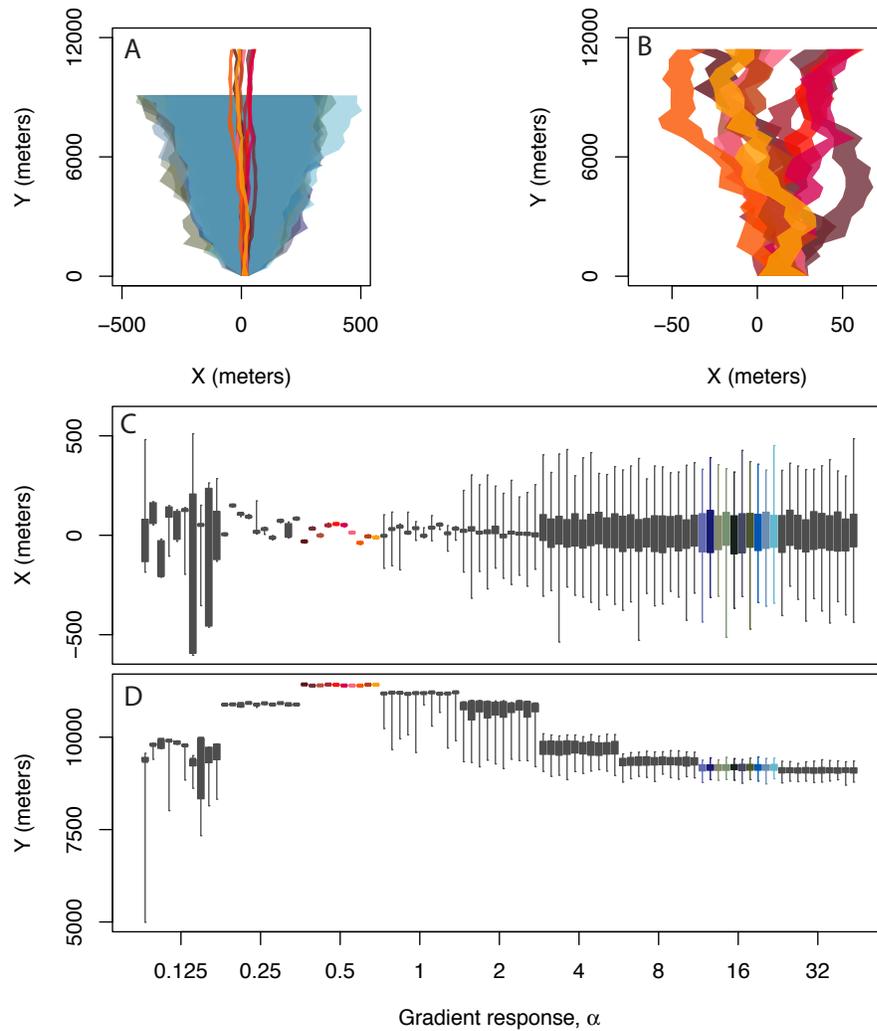


Figure 2.2: **Contrasting collective intelligence with the aggregated responses of independent individuals using ecological data.** **A** and **B**: Distribution over space and time for model individuals released at the origin and heading upwards in a noisy environmental gradient. The cool colors (blues and greens) show 10 replicate model runs for nearly independent gradient followers ($\alpha = 32$). The hot colors show 10 replicate runs for populations that exhibit collective behavior ($\alpha = 0.5$). **C**: Distribution perpendicular to the gradient direction at the end of four hours for varying levels of collective behavior, with each replicate for a given value of α shown side-by-side. Boxes enclose the interquartile range and lines enclose the entire range of the data. Colors correspond to those in the previous panes. **D**: Distribution parallel to the gradient direction.

when the underlying behaviors are not observed. Group idiosyncratic deviations from the mean preferred direction generate significantly more order in observed velocities at a particular time, compared with average velocity over time, leading to increased values for $(\bar{\psi})$ in populations influenced by collective behavior (Figure 2.3). The collective mistakes that produce this difference are the result of patterns in which a large portion of the population travels in a certain common direction at a particular time, but where that direction varies randomly over time. These dynamics are highly unlikely for populations of independent gradient followers, where independent trajectories, by definition, are as likely to show similarity within a given sampling period as among sampling periods.

The adjusted order parameter $\bar{\psi}$ is correlated with the gradient response parameter α ($R^2 = 0.49$, $p < 0.0001$, for linear regression of $\bar{\psi}$ as a function of $\log \alpha$, with each observation time in each replicate as a single data point; $R^2 = 0.5$, $p < 0.0001$ on average when the analysis is done on a single replicate). While the value of $\bar{\psi}$ fluctuates over time points and replicates, populations with the strongest collective behavior ($\alpha = 0.125 - 0.5$) are clearly distinguishable from those with low levels of collective behavior ($\alpha = 8 - 32$) based only on values of $\bar{\psi}$ (Figure 2.3A,B). In populations with intermediate levels of collective behavior, where the influence of the environmental gradient on individual velocities is at least as strong as that of social interactions, but not overwhelming ($\alpha = 1 - 4$), $\bar{\psi}$ attains intermediate values (Figure 2.3A). In some of these intermediate cases, the value of $\bar{\psi}$ varies systematically over time, due to long transient patterns caused by the aggregation of the population in the initial conditions (Figure 2.3C).

Ecological studies may not have data for high and low levels of collective behavior with which to make direct comparisons. Therefore we also compare the value of $\bar{\psi}$ to that of a null statistic $\bar{\psi}_0$ which is obtained by calculating $\bar{\psi}$ on data where the observation times t_j have been placed in a random order relative to the velocities. This procedure breaks correlations in velocities among individuals observed at the same time, and estimates the value of $\bar{\psi}$ if individual velocities were independent, using the distribution of observed velocities as a starting point. That is, a test for collective behavior in ecological data on the distribution patterns over time of a single population can consist of not just the qualitative check for higher values of $\bar{\psi}$, but a quantitative test of the hypothesis that on average over time $\bar{\psi} > \bar{\psi}_0$ (see Figure 2.3A).

We measured the power of this test as follows. First, for each replicate simulation and level of α we computed the average value of $\bar{\psi}$ over all individuals and times. Next, we repeatedly calculated $\bar{\psi}_0$ in each of $M = 100$ randomizations of the sampling times. These were also averaged over individuals and times, yielding M comparison values for each value of the average $\bar{\psi}$ calculated in the first step. For each replicate simulation and level of α we then compared the average value of $\bar{\psi}$ to the 95th percentile of the comparison values, recording for each level of α the frequency with which the average value of $\bar{\psi}$ exceeded the 95th percentile of the distribution of $\bar{\psi}_0$ s over replicates. In all but one case ($\alpha = 32$) the frequency was 1.0, indicating the test has good power, even at weak levels of collective behavior. For $\alpha = 32$ the frequency was 0.8.

We tested the robustness of our approach to unobserved individuals by repeating the analysis on randomly selected fractions of the simulation data (Figure 2.4). Downsampling causes a modest decrease in the specificity of our test

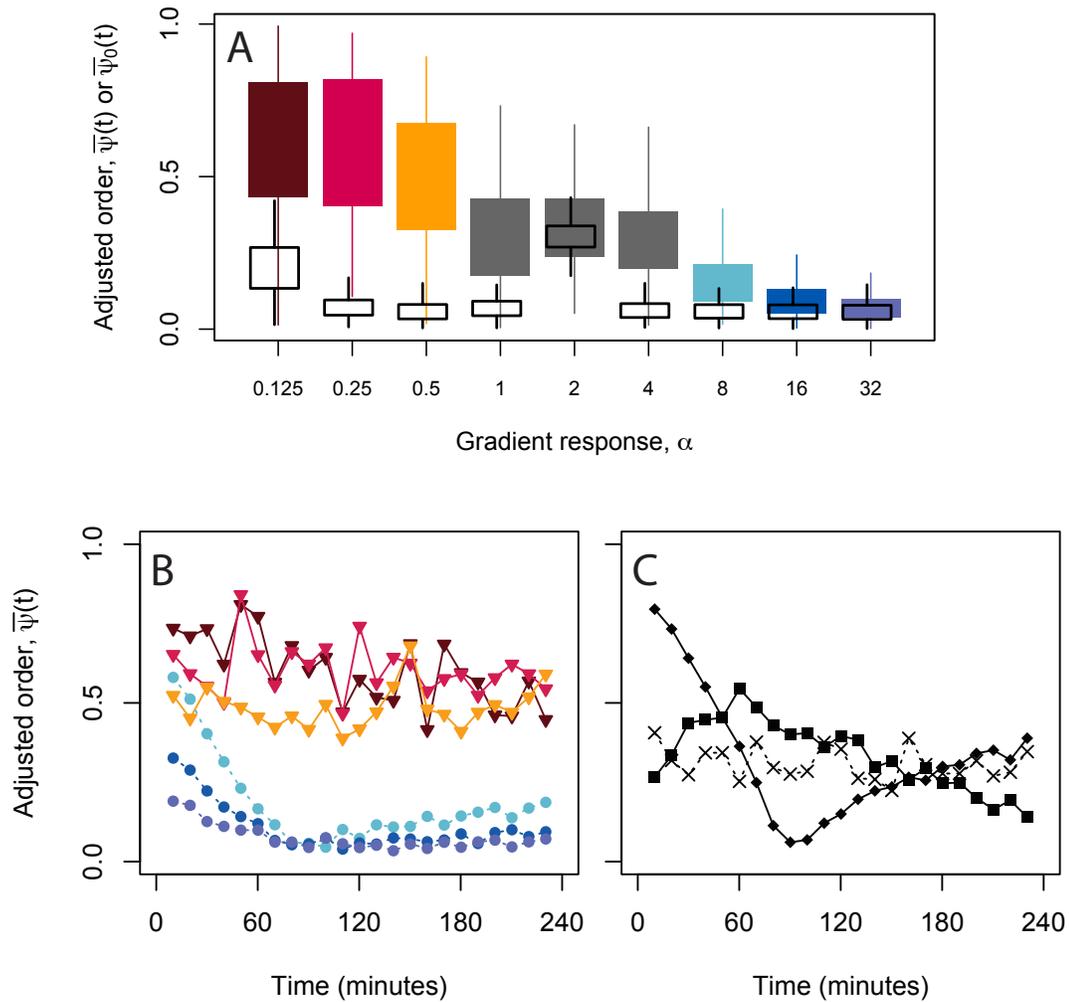


Figure 2.3: **Detecting collective behavior in ecological data.** **A:** Adjusted order ($\bar{\psi}$) in the spatial distribution of simulated populations with different levels of collective behavior. Filled boxes enclose the interquartile ranges for the distributions of $\bar{\psi}$ across all sampling times and replicates. Lines extend from the 5th to 95th percentiles of the distributions. Hollow boxes show the analogous distribution for the null parameter $\bar{\psi}_0$. **B:** Adjusted order over time for populations strongly influenced by collective behavior (α from 0.125 to 0.5; triangles) and weakly influenced by collective behavior (α from 8 to 32; circles), averaged over replicates. Colors correspond to those A. **C:** Average adjusted order over time for populations with intermediate levels of collective behavior ($\alpha = 1$, dotted line and crosses; $\alpha = 2$, solid line and squares; $\alpha = 4$, solid line and diamonds).

for collective behavior (decreasing the probability that independent populations are correctly identified) but does not significantly affect the sensitivity of the test (the rate at which populations with collective behavior are correctly identified). More specifically, sparsely sampled populations of nearly independent individuals ($\alpha = 32$) show higher adjusted order due to sampling effects, as the sample size becomes small, increasing the chances that they could be falsely identified as strongly influenced by collective behavior. However, the decrease in the specificity of our test due to unobserved individuals is modest. For instance, in populations of nearly independent individuals, mean adjusted order remains below 0.5 even if only a few individuals are observed on average per time point. By contrast, in populations exhibiting collective behavior, adjusted order remains higher than for populations of independent individuals, even if only a few individual in the population are observed.

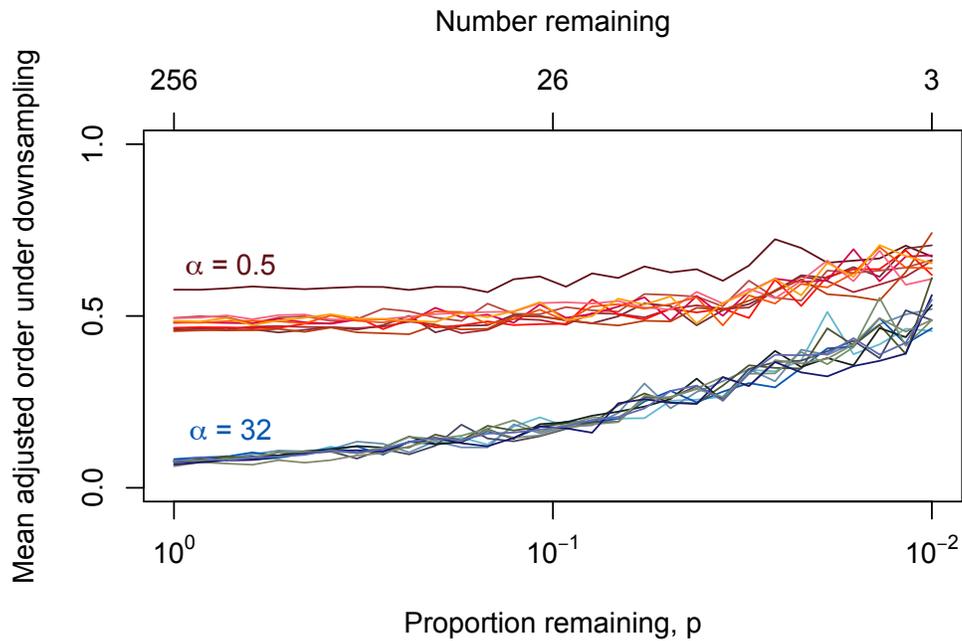


Figure 2.4: **Behavior of the adjusted order parameter when some individuals are unobserved.** Lines show the average value of $\bar{\psi}$ across replicates for populations showing collective behavior ($\alpha = 0.5$; warm colors) and populations of nearly independent gradient followers ($\alpha = 32$; cool colors), when a proportion of the individuals are randomly removed from the analysis. Each of the 10 lines of a given color and style corresponds to a repetition of the analysis on a different replicate simulation. Each line is composed of 30 points, which are each derived from a different random sample of the original simulation output.

2.4 Discussion

Collective decisions in mobile animal groups emerge by the propagation of local behavioral interactions to influence population dynamics. As such, the approach we propose for detecting collective movement in ecological data—using broad scale collective deviations from the mean gradient direction that would be unlikely for independent individuals—rests on a fundamental property of collective behavior [33, 36]. Our contribution is to identify specific quantitative features of this process that are readily observable in ecological data, where only a fraction of individuals are observed and the time between observations is much longer than the time scale of individual behavioral decisions. Without recourse to fine-scale observations of individual behavior, the approach we describe can, under some conditions, reject the null hypothesis that the data were generated by independent responses to a common environment (such as a chemical gradient), or to physiological stimuli operating independently among individuals (such as physiological responses to photoperiod).

Ultimately research on the causes and consequences of collective behavior requires identifying the underlying mechanisms that drive its emergence, maintenance and dynamics. However discovering the ecology of collective behavior in nature also requires methods for learning about its prevalence in populations that are not exhaustively sampled at the resolution of individual behavior. As in the study of ecological competition, or evolutionary adaptation, pattern-oriented “top-down” approaches to studying collective behavior can complement “bottom-up” mechanistic approaches, and the most exciting discoveries often involve a combination of both [35]. To complement high-throughput ethological approaches in laboratory and wild populations, our approach has the

power to screen for collective behavior where fine-scale behavioral data have not yet been collected, with the potential to diversify and enlarge the set of populations where collective behavior is considered.

The method we propose uses the same intuition as [56] by seeking patterns in coarser-scale data that would be unlikely for independent individuals. However, unlike [56], our approach does not require an explicit null model for the behavior of independent individuals. In our approach, v_0 plays the role of a null model, representing the expected velocity of independent individuals. Moreover, instead of specifying v_0 *a priori*, we calculate it from the data.

In many applications the preferred direction of travel ϕ will vary over space and time. In these cases it will be necessary to learn how ϕ changes and incorporate that into estimates of $v_0(t, s)$. This could be done by straightforward estimates of the response of independent individuals to the state of the environment (for example, connecting gradients in light levels to the swimming speed of fish [34], or to the velocity of phytoplankton [58]). Figure 2.5 shows an example of one such analysis, where a series of two identical experiments are used. Each experiment consists of running the collective behavior simulation with a gradient direction that changes deterministically over time. Data from the first experiment is used to estimate the gradient direction over time, while data from the second experiment is used to test for the strength of collective behavior. From the first experiment we estimate the gradient direction over time by fitting a cubic spline to the velocity data (Figure 2.5A). We then apply our approach for detecting collective behavior to the data from the second experiment using the smooth function from the first experiment as $v_0(t, s)$ (Figure 2.5B). As above, this approach uses small but significant collective deviations from the true gradi-

ent directions (which is estimated independently) to detect collective behavior (Figure 2.5C). In a more advanced approach, one set of observations might suffice for both detecting the gradient and analyzing collective deviations from it. However this would be more complicated: within a single replicate, changes in the gradient direction over time are confounded with idiosyncratic deviations from the true gradient direction due to collective behavior.

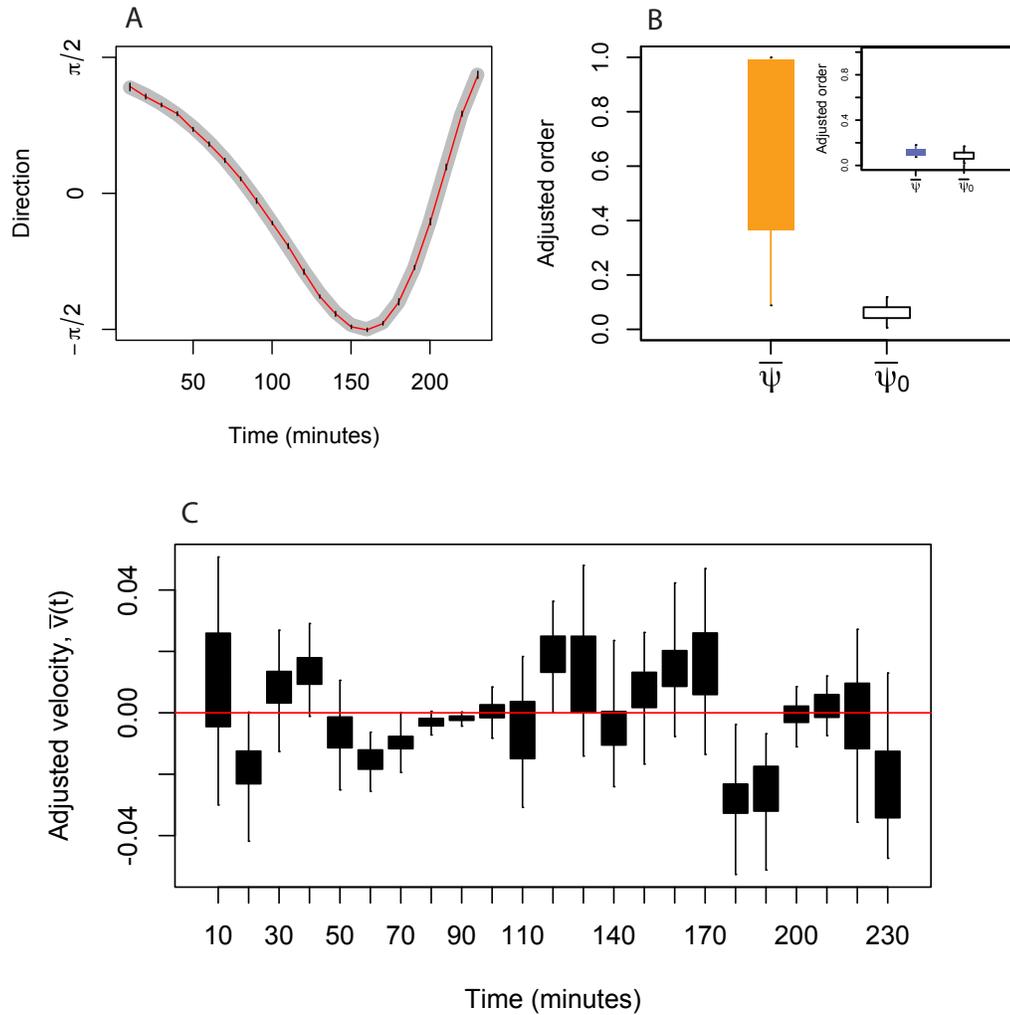


Figure 2.5: **Detecting collective decision making in a temporally shifting environmental gradient.** **A:** Direction of travel over time for a population with $\alpha = 0.5$. Vertical lines encompass the range of directions of travel for individuals observed at a particular time. The thick grey curves shows the true direction of the gradient. The red line shows a cubic spline fit to the velocity data. **B:** Adjusted order and null order over time in a second replicate realization of the process, where velocities were adjusted by subtracting the spline fit to the velocity data in the first replicate (as shown in **A**). The inset in **B** shows the analogous results for a population of nearly independent gradient followers ($\alpha = 32$). **C** shows the distribution of the horizontal component of the adjusted velocities in the second experiment, showing collective deviations from the true gradient direction, as estimated in the first experiment.

Applications for our approach include understanding the emergence and stability of seasonal migration patterns, when migration originates from a mixture of environmental/physiological stimuli and collective behavior. For example, the migration patterns of herring (*Clupea harengus*) may depend on ocean currents and food availability [59] as well as on juveniles learning the migration routes by following older age classes [60]. In humans, daily movement patterns in cities show the influence of the external (built) environment as well as the effect of social processes, such as when a significant fraction of people commute to work in a few specialized areas of the city [61, 62, 63]. So the dynamics of humans, herring, and many other populations, may depend on the interplay between individual and collective decisions. While a “bottom-up” approach to detecting collective behavior in these populations is limited by the availability of fine scale behavioral data, the signal of collective decision making is, under some conditions, detectable in coarser scale ecological data.

2.5 Acknowledgements

This chapter is based on the manuscript “Detecting collective decision making in ecological data on mobile animal groups” by Benjamin D. Dalziel and Stephen P. Ellner. BDD and SPE conceived of the experiment and did the analysis. BDD and SPE designed the analysis. BDD and SPE wrote the manuscript.

CHAPTER 3
THE DYNAMICS OF COLLECTIVE BEHAVIOR IN CARIBOU
MIGRATION

3.1 Introduction

Migration is a behavioral adaptation to variable environments that is found nearly everywhere in the tree of life. Bacteria [64], fungi [31], insects [65], birds [66], ungulates [67], and members of almost every other major evolutionary lineage migrate. Migrating individuals often face considerable risk, and spend a significant portion of their energy, in order to travel toward resources that are located outside their perceptual range, effectively using broad-scale variation in resources to compensate for local variability. One of the fascinating things about migration is how often it has evolved, been lost, and evolved again within independent lineages. This indicates that although migration can be risky and costly, it is an evolutionarily very labile trait, suggesting it derives from fundamental features of life [68]. But how do migration patterns emerge from basic characteristics of individual behavior?

Two broad classes of processes explain migration patterns. First, migration can be driven by responses to physiological or environmental cues that are independently replicated among individuals. Even when raised in isolation and deprived of migratory cues from the environment (such as photoperiod), some animals adapted to migration will display precisely-timed physiological responses (such as migratory restlessness), confirming a genetic basis for migration [69]. And there are a host of adaptations that allow animals to navigate in response to environmental cues such as magnetic fields [70], sun azimuths [71], wind [72],

and gradients in food quality [67]. However, while physiological and environmental cues are certainly important, it is unclear if they can account for all the specificity and variability of migration patterns over space, time, and taxa [68].

The second process that can explain migration patterns is collective behavior, where organisms adjust their velocities in response to others nearby, causing the formation of groups that more effectively track weak environmental signals [30, 73, 74]. Group decision making is common in animals [75], and exchanging information among conspecifics can improve navigation [39] and influence the timing of migration [76]. Observations of the results of collective behavior in animal populations include wave-like fronts in wildebeest herds [77], fission-fusion dynamics in elk populations [78], the emergence of spontaneous order in marching locusts [79, 80], quorum decision making in fish schools [81], and the hierarchical geometry of pigeon flocks [46].

Simulations agree that collective intelligence plays an important role in organizing a wide array of biological and social systems, and that it can arise from basic sensory and cognitive systems [33, 30, 34]. These simulations further suggest that populations of individuals who use collective behavior to migrate in noisy environments may easily arise in different areas of the tree of life, and have increased survival and reproduction, relative to sympatric populations who do not exhibit collective behavior [73, 74]. Collective behavior is thus an important candidate for explaining the widespread evolution of migration. However this hypothesis has rarely been tested in field data.

If collective behavior drives the evolution and maintenance of migration patterns then the statistical signature of collective behavior should be detectable in relocation data on migrating organisms. Methods for detecting collective be-

havior in animal relocation data have been recently proposed [46, 50, 47, 48], but observing a whole population at individual-level resolution (the ideal for detecting collective behavior) remains a significant technological challenge. Another challenge is distinguishing patterns that are associated with collective behavior from those arising from aggregated independent responses to a shared environment [55]. Thus while there is ample evidence for the potential importance of collective behavior in simulations and laboratory experiments, detecting the signature of collective behavior in populations of wild animals moving in heterogeneous environments remains an important and open problem [82].

Here we use long-term data on the relocation patterns of migratory caribou (*Rangifer tarandus*) to distinguish collective behavior from independent responses to a seasonally and spatially variable environment. We statistically estimate a seasonally and spatially varying velocity field using independent data, which we interpret as representing physiological and environmental cues available to all individuals. Simulations of individuals that are assumed to respond individually and independently to the velocity field reproduce very accurately the observed migration patterns of the population. However, zooming in on nearby caribou reveals that their velocities are significantly more ordered than predicted by the velocity field under the hypothesis of independence among individuals. As a result, the accuracy with which individual caribou velocities can be predicted is more than doubled if the velocities of nearby caribou are also taken into account. Finally, we find evidence that the relative importance of collective behavior varies seasonally, possibly associated with the timing of reproduction. Our study is among the first to detect the dynamics of collective behavior in a wild population of migrating organisms.

3.2 Material and Methods

Data

We study the migration patterns of the Rivière-aux-Feuilles caribou herd in Northern Québec, Canada [83]. The herd has varied in size from approximately 56,000 individuals in 1975 to at least 628,000 in 2001, to approximately 430,000 in 2011 [83, 84, 85]. These caribou usually overwinter in the boreal forest in the southern Ungava peninsula. Each spring they migrate up to approximately 1200 km to calving grounds located on the northern part of the peninsula, in tundra (61 N, 74 W; Fig. 1). Tundra is a highly seasonal environment and the arrival on the calving ground is synchronized with the peak of productivity of the vegetation at the onset of the short growing season [86]. Almost all females return to the same calving ground each year [87].

The data consist of 14,468 observations of the locations of 170 caribou observed over nine years (2003-2011). Caribou were captured using net guns fired from a helicopter, and handled without chemical immobilization. The data were collected using Argos tracking collars (Service Argos Inc., Largo, MD) that record the locations of animals every five days ($120 \text{ h} \pm 1.66 \text{ s.d.}$). The median observation period for a single animal in the data is 320 days, with approximately 30 unique individuals observed on average during any month. The data represent an unbiased subsample of the movement patterns of the herd that is small relative to the size of the herd, but large relative to most empirical studies of animal movement patterns to date. Below we describe a robust statistical approach for detecting the signature of collective behavior in a sample of data such as these.

3.2.1 Analysis

Let x_i be a two-element vector representing the i th caribou location in the data, measured as displacement (km) east and north of the southwest corner of the study area. Let t_i represent the corresponding observation time measured as hours since the beginning of the study. Define the velocity for observation i as $v_i = \Delta x_i / \Delta t_i$, where Δx_i and Δt_i are the change in location and time elapsed since the most recent prior observation of that individual. Δt varied relatively little compared to Δx and there is no correlation between Δt and v in our data, and so we treat all velocities as measured at the same temporal scale.

Our analysis posits that caribou move in a spatially and temporally variable velocity field that is generated by seasonally-varying physiological cues, seasonally and spatially varying environmental cues, and by the behavior of nearby conspecifics. This field represents the expected velocity of a caribou at a given place and time, as $\hat{v}_i = E[v_i | \phi]$ where \hat{v}_i is the expected velocity given ϕ , the parameterization of the velocity field.

We separate the velocity field into two components. First, there is the velocity generated by seasonal physiological and environmental cues enacted independently among individuals, which we represent by \tilde{v}_i . By seasonal variation we mean variation over time modulo year. Second, there is the velocity resulting from interactions with nearby individuals, \hat{v}_i . So we have

$$\hat{v}_i = \alpha \tilde{v}_i + (1 - \alpha) \hat{v}_i \quad (3.1)$$

where $0 \leq \alpha \leq 1$ is the weight given to independent environmental and physiological cues.

When $\alpha < 1$ collective behavior plays a role in determining caribou migra-

tion patterns and the resulting population redistribution patterns show complex order far from thermodynamic equilibrium [26, 88]. However, when $\alpha = 1$, movement is independent among individuals. When each caribou's position changes independently, we assume each follows the stochastic differential equation

$$dx = \tilde{v}(t)dt + \sigma dW \quad (3.2)$$

where dx represents a very small change in position, dt a very small increment of time, and σdW is white noise with total power σ^2 . The discrete time Euler approximation of this is

$$x(t+h) = x(t) + h\tilde{v}(t) + \sqrt{h}z(t) \quad (3.3)$$

which becomes increasingly accurate with increasingly small time steps, h . $z(t)$ is then a draw from a two-dimensional Gaussian distribution with mean 0 and variance-covariance matrix equal to $\sigma^2 I$, where I is the two-by-two identity matrix. The two-element vector σ^2 controls then the magnitude of the 'error' associated with a given parameterization of the field. The error represents the results of imperfections in the model, and behavioral variation in velocity among caribou at the same location and time.

Equation (3.3) produces advection-diffusion dynamics at the population level, regardless of the parameterization of the velocity field. That is, a (sufficiently large) group of independent particles moving in any velocity field will have their average velocity determined by the field (with the average being taken over the locations of the group) and the variance in their velocities determined by the ratio of signal to noise in the velocity field. Crucially, while independence among individuals leads to advection-diffusion dynamics, collective behavior does not—individuals that move collectively choose their ve-

locities based on the velocities of their neighbors, as well as being affected by the velocity field they are each exposed to.

The framework of Equations 3.2 and 3.3 encompasses many models of animal movement, including random walks, correlated random walks, and biased random walks [57]. Correlated random walks converge towards advection-diffusion on time scales that are long compared to the autocorrelation time of sequential steps[57, 89]. There are a few important exceptions to this framework, including Lévy flights [90], and state-based behavioral switching models [91]. However, populations of individuals modeled using state-based models usually also conform to advection-diffusion dynamics over long time scales.

We model \tilde{v}_i and \hat{v}_i using a kernel smoothing approach [92] so \tilde{v}_i and \hat{v}_i are the weighted average of the velocities of other caribou in respective spatial-temporal neighborhoods $\tilde{\Omega}_i$ and $\hat{\Omega}_i$ centered at the focal observation at (x_i, t_i) . In general, a neighborhood Ω_i represents the subset of the data set which the kernel smoothing model uses to predict a focal observation i .

Neighborhoods are formed based on spatio-temporal proximity. The advection-diffusion neighborhood $\tilde{\Omega}_i$ views time as modulo year, and excludes caribou observed during the same year as i . $\tilde{\Omega}_i$ thus represents the velocities of caribou independent of i near a particular location and time of year. In contrast, the collective behavior neighborhood $\hat{\Omega}_i$ consists of caribou locations from the same year as i , and represents the influence of nearby conspecifics on the velocity of the focal observation.

Each of the two neighborhoods is then defined by spatial and temporal band-

widths: $(\tilde{\kappa}, \tilde{\tau})$ for $\tilde{\Omega}_i$ and $(\hat{\kappa}, \hat{\tau})$ for $\hat{\Omega}_i$, implemented through a distance function

$$d_{ij}^2 = \left(\frac{x_i - x_j}{\kappa} \right)^2 + \left(\frac{t_i - t_j}{\tau} \right)^2 \quad (3.4)$$

such that, in addition to the above-mentioned constraints, another observation j was only included in a neighborhood Ω_i if $d_{ij} \leq 1$.

Fitting the model to the caribou movement data consists of choosing the spatial and temporal bandwidths for the neighborhoods, and choosing a value for α . We do this by minimizing squared error in predicted versus observed velocity, by cross-validation on independent data (see Appendix A). As part of the test of the importance of collective behavior in caribou migration, we fit and compare two versions of the model. We fit the full model, and one without collective behavior, where α is set to one. We call this latter model the advection-diffusion model.

3.2.2 Detecting the signature of collective behavior

Let N_i be the number of animals in a neighborhood Ω_i . Note that in our analysis neighbors are individual observations within a given neighborhood, rather than the actual number of caribou nearby to a particular individual at a given time, since most of the herd was not collared. Let $u_i = v_i/|v_i|$ represent velocity normalized to have unit magnitude. Define the level of polarization order in velocity around each individual in a given neighborhood and time as

$$\psi_i = \frac{1}{N_i} \left| \sum_{j \in \Omega_i} u_j \right| \quad (3.5)$$

which ranges from 0 when the directions of the velocities are uniform random, to 1, when all velocities point in the same direction [93, 94]. Any neighborhood

Ω_i surrounding a focal measurement i has its own associated level of order ψ_i . We use $\tilde{\psi}_i$ to represent the order associated with $\tilde{\Omega}_i$, the neighborhood that estimates independent responses to spatially and seasonally varying physiological and environmental cues. $\mathring{\psi}_i$ represents the order associated with $\mathring{\Omega}_i$, the neighborhood representing collective behavior. We show below how comparing the level of order associated with different neighborhoods can allow inference into the relative importance of environmental/physiological processes versus collective behavior processes in seasonal migration patterns.

For an ensemble of unbiased independent random walkers, ψ_i varies systematically with N_i , with expected value approximately $\sqrt{\theta/N_i}$, where $\theta = 2/\pi$ (see Appendix A). Therefore in field data there is the question of what constitutes a significant amount of order. To determine the probability that an observed ψ_i could be generated by a group of N_i independent random walkers, we can use the fact that the variance of ψ about its expectation is given by $(1 - \theta)N_i^{-1}$ for a group of random walkers (see Appendix A). In the test for the dynamics of collective behavior described below, values of ψ_i greater than two standard deviations above the expectation for independent random walkers are considered to be significant evidence for collective behavior. However, the conclusions from this analysis are robust to a wide range of choices for the significance threshold, from one to five standard deviations (results not shown).

We released virtual particles into the velocity field associated with the best estimates of \tilde{v}_i , updating their locations using equation (3.3), over the same time frame as the data (see Appendix A). We then used the same approach on these particle velocities as on the actual data to calculate the order parameters $\tilde{\psi}_0$ and $\mathring{\psi}_0$. Here the ‘naught’ subscript refers to the fact that these values come from

simulations representing the null hypothesis that the observed order in caribou velocities is generated solely by a seasonally and spatially fluctuating physiological and environmental cues.

The collective behavior hypothesis predicts significant differences between the time series $\tilde{\psi}$ and $\dot{\psi}$. If caribou rely on collective behavior during migration, then zooming in on groups of caribou at a particular location and time should reveal increasing levels of order, that is $\dot{\psi} > \tilde{\psi}$. Conversely, if caribou migration patterns are generated solely by spatially and seasonally-varying environmental and physiological cues enacted independently among individuals then it is unlikely that their movements will be more highly ordered than the velocity field which independently generates each of their trajectories. Thus, for independent particles we should have $\dot{\psi} \approx \tilde{\psi}$.

3.3 Results

Caribou relocation patterns show strong biannual oscillations corresponding to the spring (northward) and fall (southward) migrations (Figure 3.3). These oscillations are cohesive and statistically stationary. That is, the interquartile range in locations within any 30 day window is much smaller than the total range of variation in the data, and the long-term average location, as well as the timing and magnitude of the peaks and troughs in location, do not change systematically over the study period. Time series of velocity modulo year show the same conserved biannual fluctuations, but velocity modulo year is more widely dispersed than the location data. Interestingly, while the migration patterns are conserved from year to year, the spatial pattern (route) of the northward migra-

tion differs systematically from the southward one.

The kernel smoothing model with collective behavior omitted ($\alpha = 1$; see Equation 3.1) explains 28% of the total variation in caribou velocity under cross-validation. Thus roughly one third of the total variation in observed caribou velocities is attributable to seasonally fluctuating physiological cues and seasonally and spatially varying environmental cues that operate independently among individuals. The best spatial and temporal bandwidths for this model are $\tilde{\kappa} = 170.5$ km and $\tilde{\tau} = 32.6$ days. The kernel bandwidths indicate that these cues operate on a spatial scale equivalent to about two degrees of latitude, and a temporal scale of approximately one month.

While this model does not explain the majority of the observed variation in caribou velocities, simulated non-interacting particles released into this model's velocity field precisely reproduce the broad-scale relocation patterns of the herd (Figure 3.3). Both the median location and the distribution of the herd over time are well-matched between the actual caribou and the independent particles released into the field. Moreover, the timing of the migration, and the distinct differences between the north-bound and south-bound migration routes are also matched by the particles. At the same time, the behavior of the particles and the caribou show clear qualitative difference, despite good performance of the model at broad scales.

Adding collective behavior to the model greatly improves its performance under cross validation: the full model explains 63% of the total variation in caribou velocity. As expected, the spatial and temporal bandwidths on the collective behavior neighborhood of the model are smaller than the physiological/environmental ones, particularly for the temporal bandwidth, at $\hat{\kappa} = 129.1$

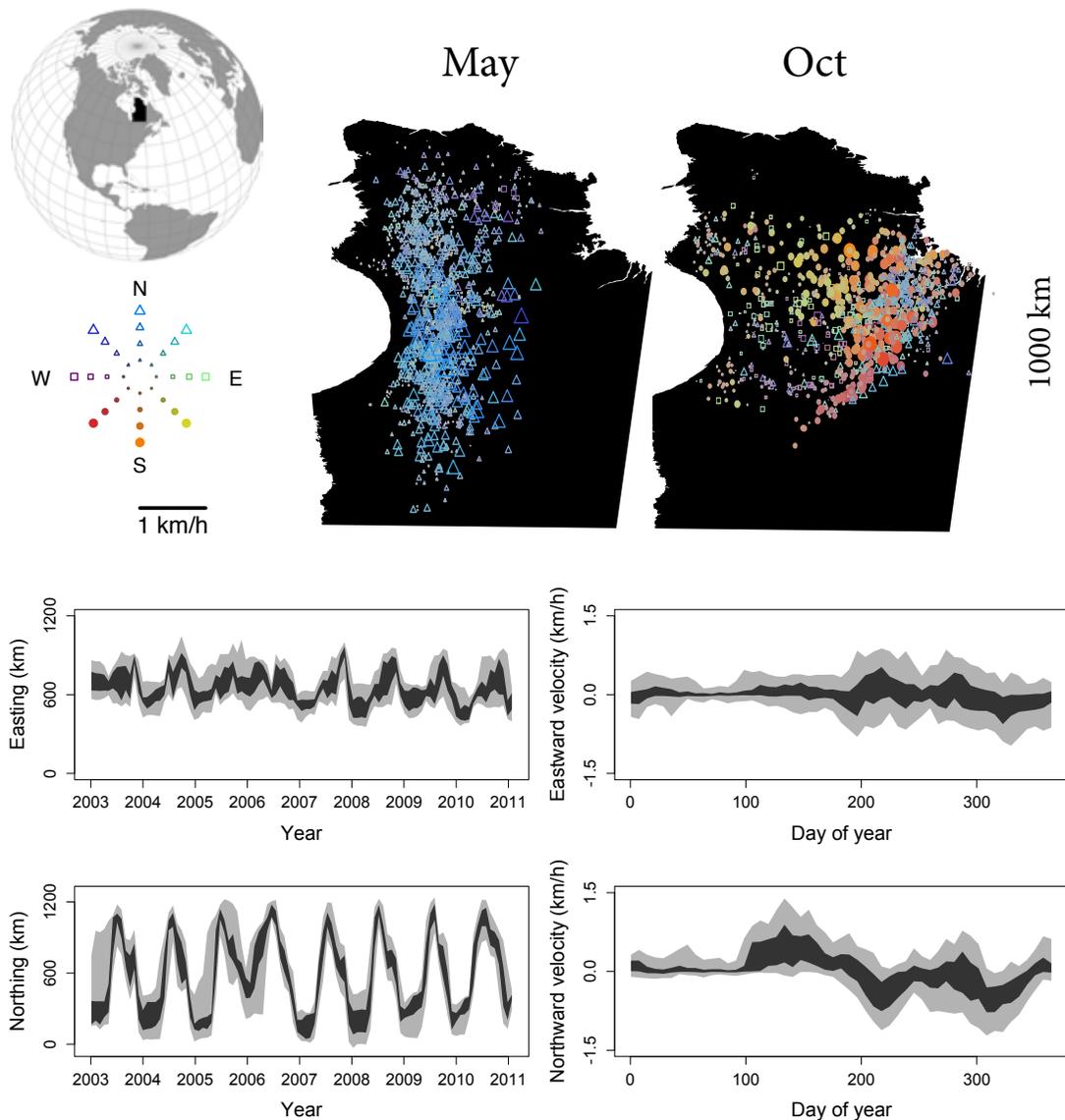


Figure 3.1: **Migration patterns of the Rivière-aux-Feuilles caribou herd.**

The inset globe shows location of the study area. The two larger maps show the locations and velocities of GPS-collared caribou observed during the spring (May) and and fall (October) migrations, pooled across years from 2003 to 2011. The style of the points show each caribou’s velocity, according to the legend on the left. Bottom panels show time series of individual locations and velocity over the study period. Dark regions enclose the interquartile range; lighter regions enclose the 5th to 95th percentiles, for running quantiles with non-overlapping adjacent windows. The window widths were 30 days for location and seven days for velocity.

km and $\tau = 6.2$ days. Collective behavior is weighted much more heavily in the full model than independent cues, with $\alpha = 0.12$. So, although a model without collective behavior can explain considerable variation in the velocity data of individuals, and can precisely reproduce the migration patterns of the herd under particle simulations, the full model gave approximately seven times more weight to collective behavior than independent behavior, and explained more than twice as much of the observed variation in velocity.

The ψ time series revealed clear seasonal fluctuations in the level of order in the herd, with spikes in order matching the timing of migration events (Figure 3.3). The simulations and the actual data matched well at the advection-diffusion bandwidth (compare the red time series in the top and bottom panels of Figure 3.3). At the same time, the probes on collective behavior reveal strong differences between the movement of the caribou and those of non-interacting particles, at the bandwidths associated with collective behavior (compare the blue time series in the top and bottom panels of Figure 3.3). The movement patterns of the caribou are significantly more ordered at these smaller bandwidths, indicating that caribou mobility patterns derive from processes other than those represented by the velocity field of the advection-diffusion model.

The dynamics of collective behavior vary seasonally in the caribou. Loosely, at some times of year, when you “zoom in” on nearby caribou, transitioning from the broader bandwidths associated with advection-diffusion to the narrower bandwidths associated with collective behavior, the level of order in their velocities does not change much. But at other times of year when you zoom in, nearby caribou are much more ordered than expected for independent particles. We measure this relative prevalence of collective behavior using the ratio

$\dot{\psi}/\tilde{\psi}$, which increases as the velocities of nearby caribou become more highly ordered, relative to the predictions of the advection diffusion model. There is a spike in the relative importance of collective behavior which occurs each year in July (Figure 3.4). This spike occurs soon after calving and just as the herd reaches its latitudinal zenith.

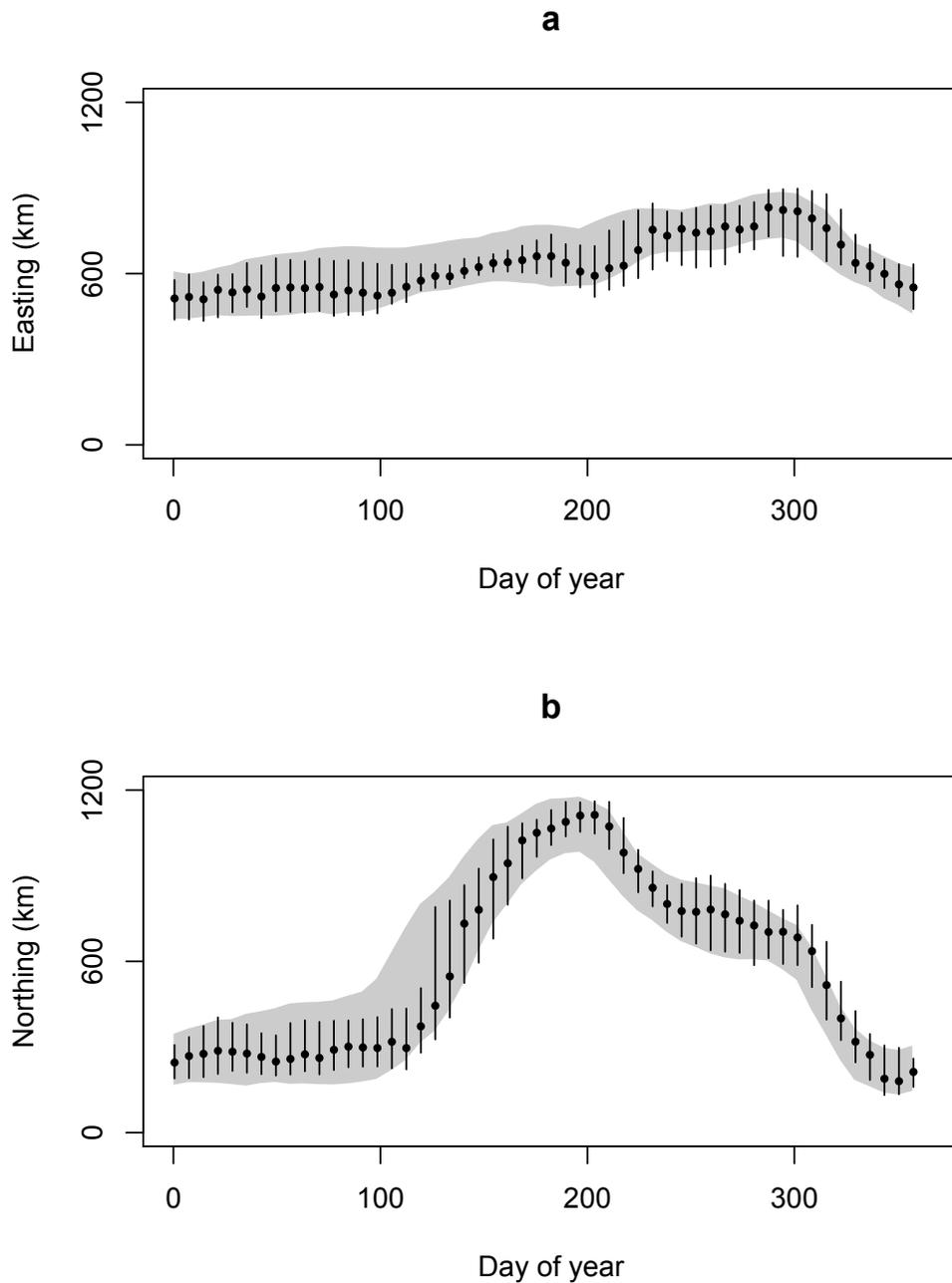


Figure 3.2: **Performance of the advection-diffusion model of caribou migration, simulated by releasing independent particles into its velocity field.** Panes show easting (a) and northing (b) as a function of time of year, for caribou (points and lines) and particles (shaded area). Points show running median caribou locations using seven day consecutive non-overlapping windows. Vertical lines enclosed the interquartile range for the same windows. The shaded region encloses the analogous interquartile range for the particles.

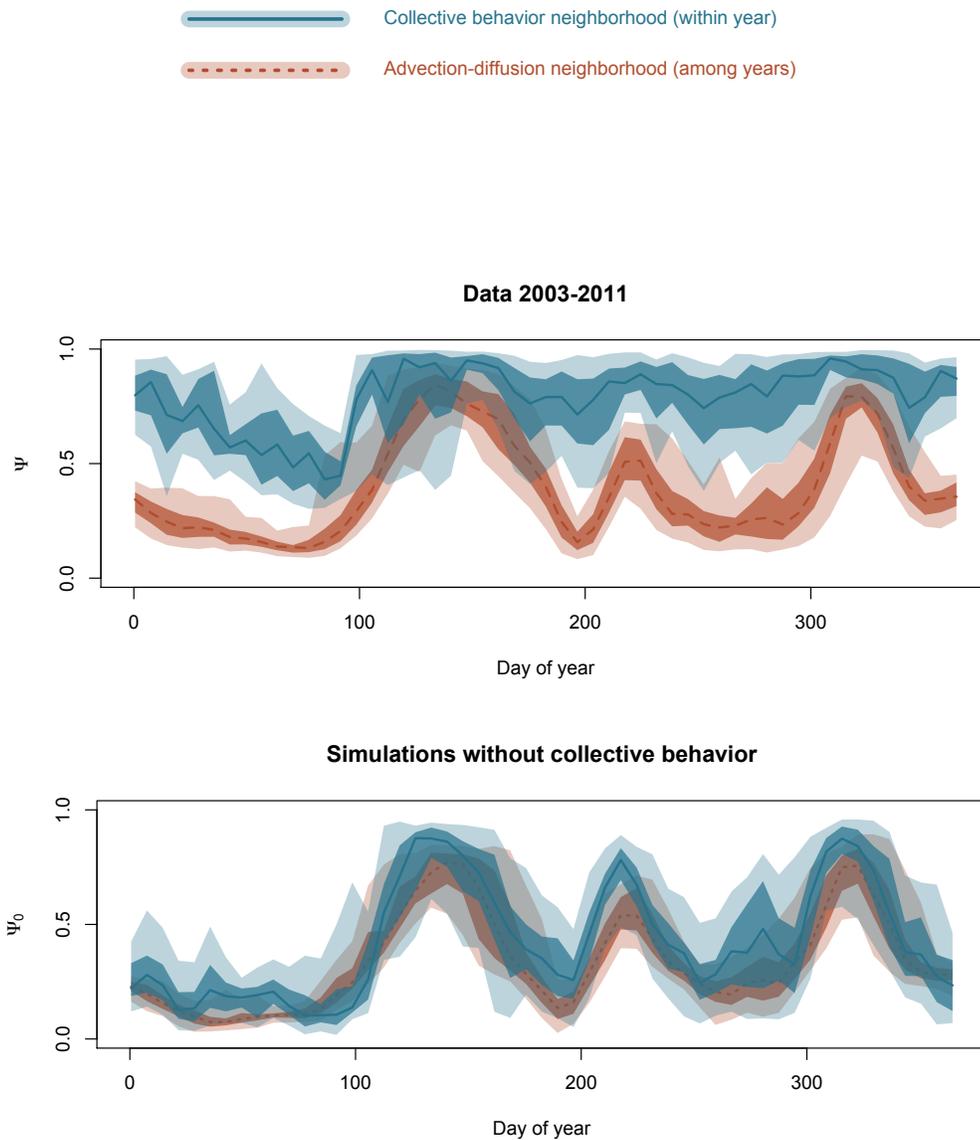


Figure 3.3: **The signature of collective behavior in caribou migration patterns.** Top panel: time series, modulo year, of the order parameter ψ for the spatio-temporal neighborhoods associated with collective behavior (blue, solid line) and advection-diffusion (red, dashed line). Lines show the medians, darker regions enclose the interquartile range, and lighter regions enclose the 5th to 95th percentiles, for running quantiles with 7 day non-overlapping adjacent windows. Bottom panel: the same analysis done on simulations where caribou do not interact.

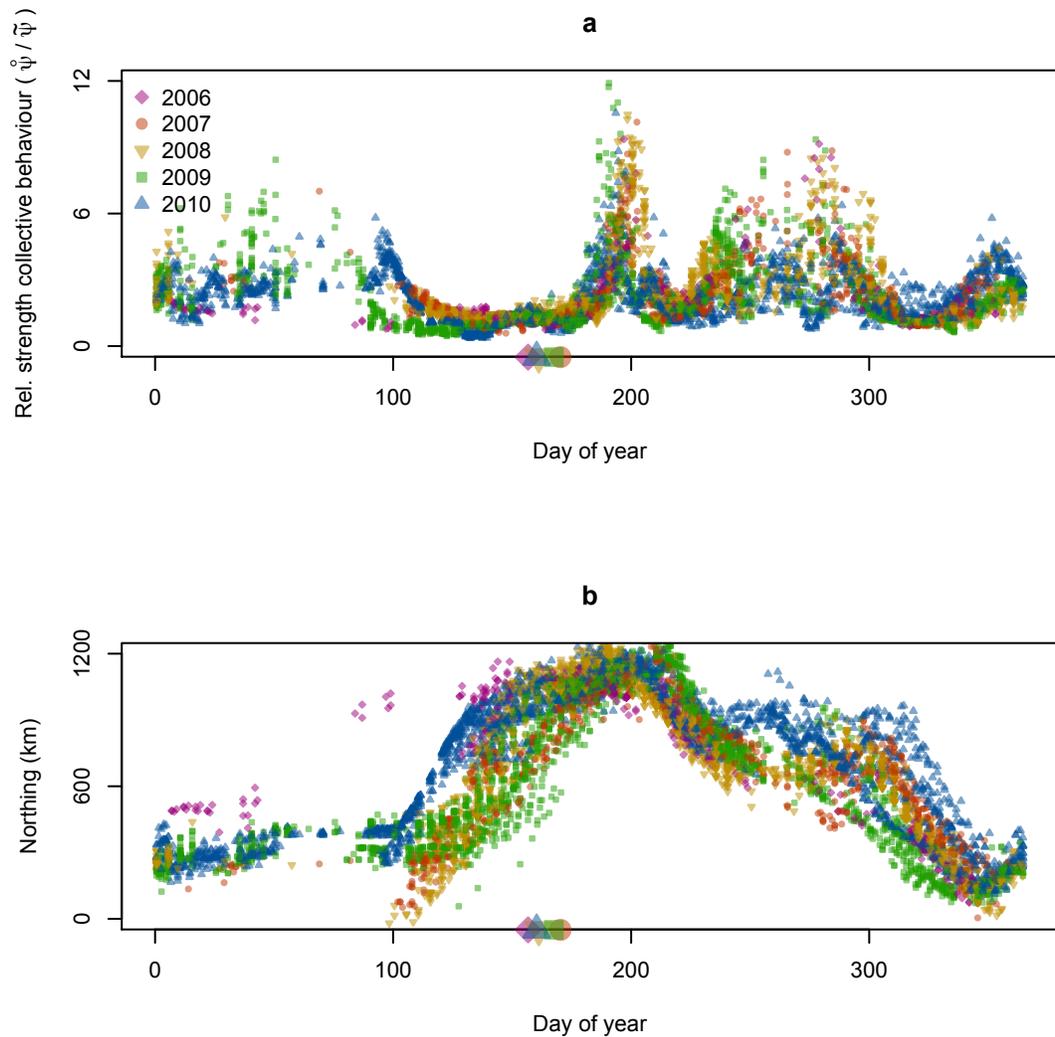


Figure 3.4: **A spike in collective behavior of migrating caribou each year in July.** (a) Points along the horizontal axis show the average date of calving (b) Latitude of the the caribou for the same time range; the spike in collective behavior seems to be right before they arrive at their farthest north location, and when herd density is high.

3.4 Discussion

The migratory caribou we study display large-scale seasonal variation in their velocities that cannot be parsimoniously explained by independent responses to seasonally fluctuating physiological cues or a seasonally and spatially fluctuating environment. Rather, in our analysis the majority of variation in the velocities of caribou is attributed to correlations among the velocities of nearby individuals, operating in addition to the physiological/environmentally-driven advection field they are each exposed to. Collective behavior may therefore play an important and dynamic role in animal migration patterns - more so than has been previously shown.

Given that result, it is surprising how well the advection diffusion model performs when we forward-simulate with independent particles. The success of the particle simulations at predicting relocation patterns shows that the advection-diffusion model is not a straw-man: the model was fit to velocity data and in the particle simulations, there is ample opportunity for errors in the model to accumulate as position integrates forward in time. Since the particles nonetheless reproduced the migration patterns (year after year), this is strong evidence that the velocity field is a good model for the advection-diffusion component of caribou migration. The poor predictive power of the model under cross validation against individual velocities shows, then, that there are processes generating order in velocities beyond what a seasonally and spatially varying advection-diffusion model can account for.

Correspondingly, we find that the velocities of nearby caribou are significantly more ordered than can be accounted for by the advection-diffusion

model. Fluctuations in the level of order within groups of nearby caribou indicate that the influence of collective behavior on caribou relocation patterns is dynamic. These fluctuations coincide with reproduction, suggesting that collective behavior is not just important for relocation patterns but can be a dynamic part of the life history of animal populations. We do not know what ecological processes cause the spike in collective behavior after calving each year—perhaps it could be related to movement to the summer grounds, where the herd is led to by certain experienced females.

Along with these results, we note that our method of partitioning variance in relocation data into independent behavior and collective behavior is far from perfect. The ‘independent responses’ described by the advection-diffusion model rest on the predictive power of relocation data collected in other years. And what we call collective behavior is the additional predictive power on caribou velocities gleaned by including relocation data of animals actually nearby, after we have predicted all we can from the independent data. Therefore, if collective behavior generates the same spatiotemporal pattern in velocities year after year then our partitioning approach would fail to detect any signal of collective behavior.

For example, preceding the spike in the relative importance of collective behavior in July is a period when the relative importance of collective behavior appears to be low, which corresponds to the spring migration. This depression originates from the fact that caribou migrations northward are so coherent from year-to-year that they are well described by a seasonal advection-diffusion model. Our test of collective behavior is therefore conservative in that it overestimates the importance of independent responses to a seasonal and spatially

varying environment, or seasonally varying physiological cues. In the case of our results, we can then confidently reject the hypothesis that such cues fully explain migration, since even our inflated estimates of the importance of independent cues were moderate.

Some of the signal we identify as collective behavior may be not driven by spontaneous order formation in groups but instead by localized (non-seasonal, fine spatial scale) environmental cues. However, at $\tau = 6.2$ days, caribou are mostly paying attention to events within a one week window, centered on the present time. Whatever they are paying attention to is changing very fast, or data from a larger time-window would be useful, and τ would be larger. This rules out many types of environmental variation as potential “local” cues, and strengthens the case that caribou are paying attention to each other’s velocities. Moreover, it is highly unlikely that an important driver of large scale seasonal migration patterns is local, nonseasonal environments.

The two classes of processes that explain migration—independent responses to physiological and environmental cues, and collective behavior—predict contrasting migration dynamics under increasing environmental variability. As the environment becomes more noisy, physiological/environmentally driven migration will eventually deteriorate as the cues for migration become more difficult to detect, the target resources are no longer available at the right place and time relative to phenology, or because habitat destruction or the creation of anthropogenic barriers alters mobility patterns [95, 86, 96]. In contrast, migration that is driven in part by collective behavior may display stability in the face of increasing environmental noise through the efficient propagation of information from a few informed individuals to the rest of the group.

Studying the interaction between independent responses in a shared environment and collective behavior furthers our understanding of how individual behavior scales up to affect population dynamics in variable environments. As environments become more unpredictable due to climate change, habitat destruction and other anthropogenic effects, the viability of migratory populations will depend on how collective behavior affects their spatial and temporal dynamics. Understanding the interplay between independent and collective behavior in migration is thus important for conserving biodiversity.

An interesting direction for future theoretical work would be to develop particle simulation parameterized for caribou that exhibit collective behavior. However, when the magnitude of velocity is variable, these models require a statistical function that governs the speed of highly polarized groups [97]. Estimating such a function from data would not change our results, and is outside the scope of this paper. We wonder though if the instability in variable-velocity models of collective behavior might be biologically important. Collective behavior is, in some ways, an amplifier. It might be therefore adaptive to have a combination of collective behavior and response to environmental and physiological stimuli, to keep group velocities stable and tuned to the environment. Future models might also incorporate nutritional state, which interacts with collective behavior [98].

Populations are by definition composed of individuals who can interact, and aggregation and information sharing among nearby organisms is a common feature of life. Yet collective behavior is rarely included in models of spatial or temporal population dynamics. Under what conditions does collective behavior significantly affect ecological and evolutionary dynamics in wild populations?

3.5 Acknowledgements

This chapter is based on the manuscript “The dynamics of collective behavior in caribou migration patterns” by Benjamin D. Dalziel, Mael Le Corre, Steeve Côté and Stephen P. Ellner. BDD and SPE conceived of the study. BDD and SPE designed the analysis. BDD performed the analysis. MLC and SC collected the data. BDD, MLC, SC, and SPE wrote the manuscript.

CHAPTER 4
HUMAN MOBILITY PATTERNS PREDICT DIVERGENT EPIDEMIC
DYNAMICS AMONG CITIES

4.1 Introduction

Infectious diseases cause morbidity in most humans each year [99], and account for a significant portion of all yearly human mortality [100]. As the global urbanization rate continues to rise past 50%, cities will more often act as focal points for epidemics: providing venues where strangers are more likely to interact, representing the most probable locations of first detection, and incurring a greater share of the casualties. Given cities' pivotal role in the spread of infectious disease, it is important to understand why they exhibit systematic variation in the timing and severity of epidemics [101, 28, 102].

Human mobility patterns generate the proximity between individuals prerequisite for the transmission of many infectious diseases. This suggests that cities with different mobility patterns may also differ in the rate at which their inhabitants have infectious contact, leading to variation among cities in the risk of an epidemic [102, 103, 104]. Human movement patterns are heterogeneous at a wide range of scales—from within a building [105] to among countries [106, 107, 108], as evidenced by diverse sources of data, including the movements of cell phone users [109, 110], air travel patterns [106, 107, 108] and census data on commuting patterns [111, 107, 112]. At each scale, there appear collective mobility patterns maintained far from those predicted by homogeneous random movement. These dissipative structures [26] have the potential to create localized areas where infectious contact rates are systematically amplified

[113]. Empirically reconstructed contact networks suggest systematic variation in infectious contact rates across countries, age groups, and other sociodemographic factors [114].

Individual variation in rates of infectious contact can significantly alter patterns of disease spread [22, 23, 24, 104, 103, 112] and theoretical models of disease dynamics within and among cities (both individual-based simulations [115, 102, 107, 104, 112, 116, 117] and metapopulation models [111, 118, 119, 120, 107, 108, 112]) have shown that heterogeneous contact patterns are potentially important in determining urban epidemic dynamics. However, few studies have examined whether empirical variation in intra-city mobility patterns is sufficient to drive detectable differences in epidemic dynamics among cities.

Here we use 2006 census data on the mobility patterns of 7,225,810 workers in 48 cities to test whether cities differ enough in their mobility patterns to generate differences in their risk of an epidemic. In the first part of the paper we quantify differences in mobility patterns among cities, using heterogeneity statistics and transportation models to examine whether cities vary systematically in the level of organization in their mobility patterns. In the second part of the paper we use the commuting data to parameterize a basic model for the spread of an airborne pathogen in each respective city, to test whether the observed differences among cities in mobility patterns are sufficient to generate significant differences among cities in the risk of an epidemic.

4.2 Methods

We analyzed data from the 2006 Canadian census [121] on the commuting patterns of every worker in 48 Canadian cities—a total of 7,225,810 individuals (Table 1 in Appendix B). The data for each city are organized by census tract (CT) and record the number of workers who live in CT i and work in CT j , denoted T_{ij} . CTs are contiguous geographic areas where 2500-8000 people reside and are typically the same area as a few city blocks. While the geographic area of a CT is influenced by residential population density, the boundaries of CT are chosen without regard to the number of individuals who travel to work there [122]. Let $n_i = \sum_j T_{ij}$ represent the number of workers who reside in CT i . Let $m_j = \sum_i T_{ij}$ represent the number of workers who travel to work in CT j . Let \bar{m} and σ_m represent the mean and standard deviation of m_j in a given city. Let $W = \sum_i \sum_j T_{ij}$ represent the total number of workers in a city and N the total population of the city across all CTs that have any workers. Lloyd’s “mean crowding” statistic [123] $m^* = \bar{m} + \sigma_m^2/\bar{m} - 1$ measures worker density from the perspective of workers in their workplaces: in a given city, m^* is the average number of other individuals who work in the same CT as a worker chosen at random, while \bar{m} is the average worker density in a CT chosen at random.

To further characterize inter-city differences in human mobility patterns, we compared how well the patterns for each city could be explained by alternative transportation models [124]. These models respectively describe processes that lead to different degrees of organization in human mobility. The models took the form $\langle T_{ij} \rangle = p_{ij}n_i$, where p_{ij} is the probability that an individual who resides in CT i will travel to work in CT j , and $\langle T_{ij} \rangle$ denotes the expected number of commuters between i and j under the transportation model specified by p_{ij} . We

first used a configuration model [22] for the inter-CT commuting network in each city, which makes the neutral prediction that the probability that a worker who resides in i will work in j is proportional to the total number of individual workplaces in j . The configuration model is closely related to the gravity model from transportation analysis, whose suitability for analyzing mobility data is under debate [124] (see Appendix B). The second model we examined was the newly-described radiation model [124] of transportation patterns, which modulates diffusive flow between two locations by accounting for the number of potential destinations in the area between them. Both the configuration and radiation model are parameter free.

We asked if the systematic differences in mobility patterns we discovered were sufficient to cause differences in epidemic dynamics among cities by using the commuting data to parameterize a stochastic, spatially-explicit, individual-based model of airborne pathogen transmission for each city (see Appendix B). Epidemic dynamics that result from home-work movements can also be modelled using recently-developed metapopulation frameworks [125, 120, 118, 119, 112]. These models aggregate individual behavior to consider host mobility and disease spread patterns between subpopulations. Accordingly, in the Appendix B we explore the consequences of relaxing correlations arising from the preservation of individual identities in our model.

To implement the model we first used the commuting data to estimate the frequency of contact between each possible pair of workers in a city. We then translated contact frequency into pairwise transmission hazard using a basic model of within-host pathogen dynamics for acute infections. We did this for a range of pathogen transmissibilities (here pathogen transmissibility, λ , ex-

presses the strain-specific ratio between within-host pathogen load and transmission hazard; $\log_{10}(\lambda) \in \{0, 0.25, 0.5, 0.75, 1\}$). Our model makes two assumptions: first, we assume that the spatial trajectories of humans in cities can be predicted by their home and workplace locations, which is supported by recent analyses of high-resolution data on the relocation patterns of cell phone users [110]; second, we assume that excursions from an individual's bed or work station are governed by a stochastic process that is identically distributed across cities. This leads to transmission patterns that conform to mass-action within the radius of motion of an individual, but are determined by the commuting data at the scale of a city. Although in reality cities are connected by inter-city commuting, these connections are relatively weak compared to intra-city commuting, and we are testing the prediction that cities can have differences in epidemic dynamics generated endogenously by intracity mobility patterns. Thus we model each city separately. If inter-city variability in commuting patterns is sufficient to generate differences among cities in epidemic risk, our model of transmission should predict different disease dynamics in different cities for the same pathogen transmissibility.

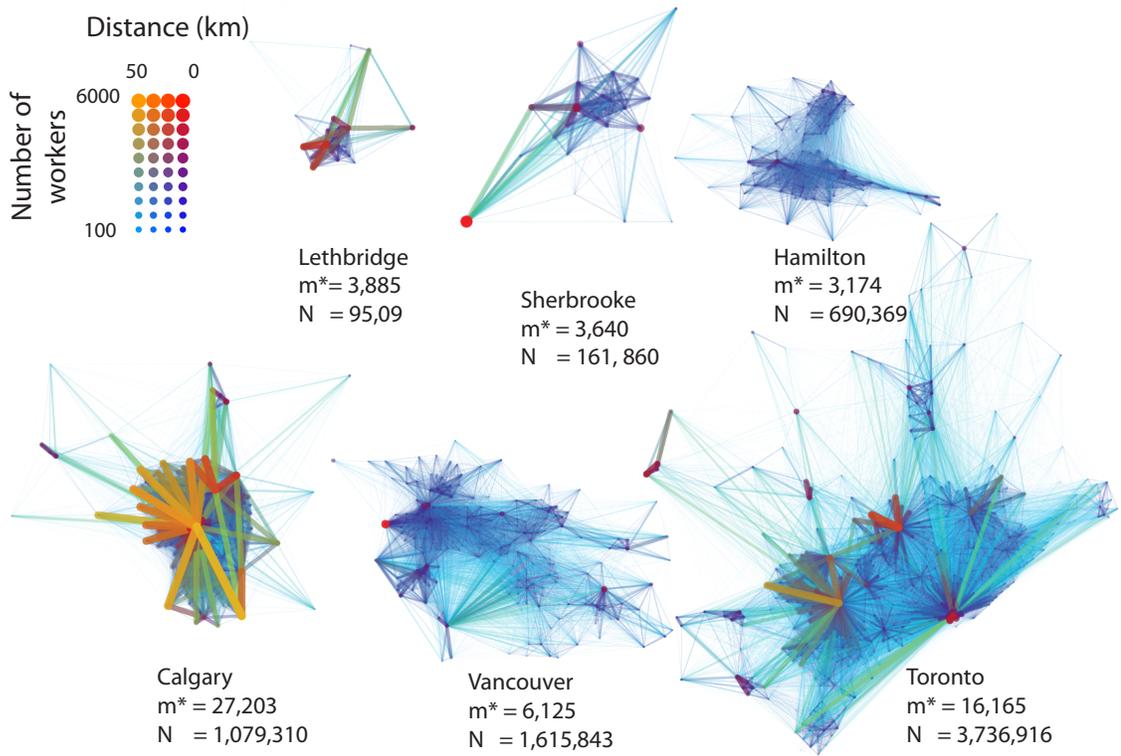


Figure 4.1: **Mobility patterns of workers in cities.** The thickness and color of edges show the number of individuals commuting between census tracts (CTs). Circles are actually short edges, representing individuals who live and work in the same CT. Larger cities tend to have more highly organized commuting patterns, as measured by the average number of workers who have their workstation in the same CT as a randomly chosen worker (m^*). However, cities also show marked differences in organization that are independent of population size

4.3 Results

The structure of the commuting matrix, T_{ij} , varies markedly both within and among cities (Figure 4.1). A striking feature in these visualizations is the appearance of star shapes in some cities. Star shapes appear where commuting flows originating from many distinct CTs are directed toward a single centralized work location. The crowding statistic m^* can be used as a measure of the prevalence of star shaped commuting patterns in a city because its value increases with the level of aggregation in individual workplace locations [123].

We find that the average number of workers per CT, \bar{m} , saturates rapidly as N increases. In contrast, m^* , which then measures overdispersion in individual workplace locations, exhibits a strong positive correlation with N . This indicates that workers in larger cities tend to organize into a few extraordinarily populated work areas, while maintaining the same average number of workstations per CT as smaller cities (Figure 4.2A). In other words, the prevalence of star shaped commuting flows in a city is only weakly correlated with the average number of people who work in a CT, but nonetheless varies strongly with total population size, leading commuting patterns in larger cities to be more highly organized around a few focal work locations. Cities also show marked size-independent variation in m^* , evident in the the ratio of m^* to the value predicted by a regression of m^* as a function N (Appendix B). In the 48 cities we analyzed, m^*/\hat{m} (hereafter “excess heterogeneity in mobility patterns”) ranged from 0.43 to 3.07, a 7.14-fold difference. For two cities chosen at random, the average ratio between the larger and smaller values of m^*/\hat{m} is 1.55. This size-independent difference in mobility patterns is equivalent to the predicted size-dependent difference (based on the regression line in figure 2a) that would result from a

2.64-fold change in population size. Thus, differences unrelated to population size are an important component of the variation in worker mobility patterns between cities.

The configuration model explains much of the variation in commuting flow in small cities, but its performance decreases systematically with N , indicating that larger cities have increasingly highly-organized mobility patterns (Figure 4.2B). The fit of the gravity model is poorer than that of the configuration model, but it exhibits the same trend of fitting smaller cities better than larger cities (see Appendix B). The radiation model also shows systematic variation in performance: as the fit of the configuration model declines, the performance of the radiation model increases, performing relatively poorly in small cities and better in larger ones (Figure 4.2B).

Cities with more organized commuting patterns (meaning a larger value of m^*) are predicted to have a higher probability (P) of an epidemic following the introduction of a single randomly-chosen infected individual (Figure 4.3A). This relationship persists once the effects of population size on m^* and P are removed: among cities of the same size, increased excess heterogeneity in mobility patterns is predicted to cause a significant increase in the risk of an epidemic, relative to the average risk for a city of that size (Figure 4.3B). The change in relative risk produced by increasing excess heterogeneity in mobility patterns is greater for pathogens with lower transmissibility. Less contagious pathogens also show more variability in relative risk among cities.

The predicted number of individuals infected by the end of an epidemic, F , also scales with the level of organization in commuting patterns (Figure 4.3C). As with the probability of an epidemic, the influence of organized host mobility

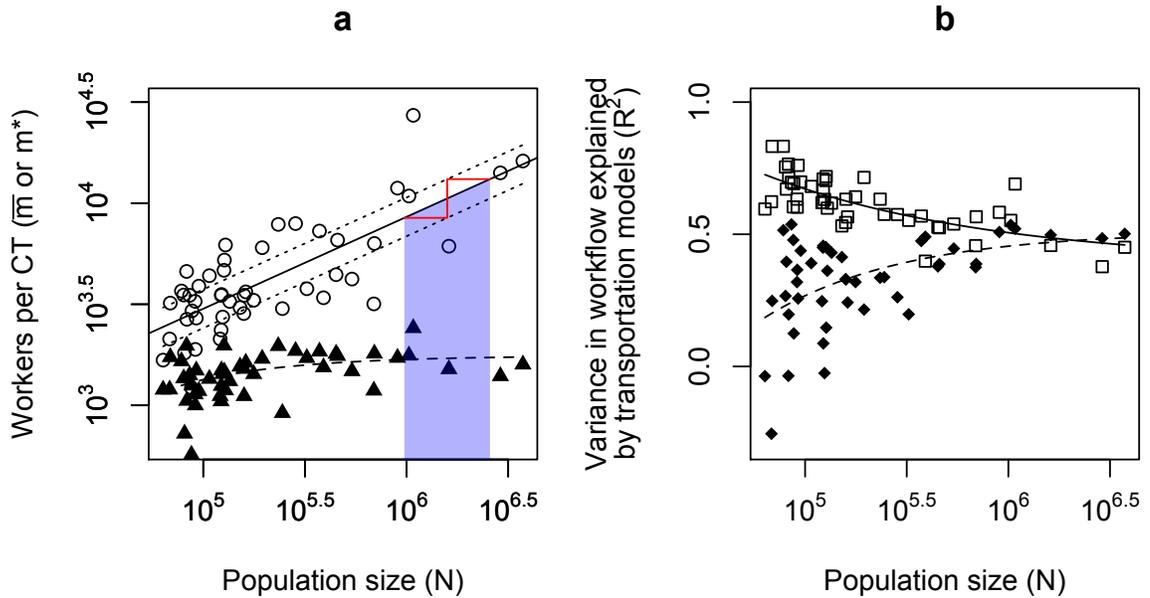


Figure 4.2: **Systematic differences in worker mobility patterns among 48 Canadian cities.** **A:** \bar{m} , the mean number of workers per census tract (CT) (triangles) and m^* , the average number of workers in the same CT as a randomly chosen worker (circles), as a function of population size (N). The solid line shows \hat{m} , the fitted relationship between m^* and N . The vertical distance between the dashed lines spans $\frac{2\sigma}{\sqrt{\pi}}$, where σ is the standard deviation of m^*/\hat{m} , showing the expected absolute difference in m^* (on a \log_{10} scale) between two cities of the same size. The width of the shaded polygon then shows what change in N would produce that difference according to \hat{m} . **B:** Variance explained in each city by the configuration (squares) and radiation (diamonds) models of commuting flows.

patterns on F is still significant once the effect of population size is removed (by considering the effect of excess heterogeneity in mobility patterns on excess infected - F/\hat{F} ; $\hat{F} = w_\lambda N^{x_\lambda}$; Figure 4.3D). In sum, the simulations show that extant differences among cities in the level of organization in human mobility patterns are sufficient to significantly alter the risk and severity of an epidemic among cities. The average magnitude and variability of this effect depends on pathogen transmissibility.

4.4 Discussion

Larger cities depend on higher levels of organization that increase economies of scale [61, 62]. Here we show that increasing organization in cities may also have important consequences for the spread of infectious disease. Whereas epidemic models have typically assumed that human populations are identically mixed for the purposes of infectious contact, our results add to an increasing body of empirical evidence that infectious contact rates in humans vary systematically among populations [126, 114, 104]. Correspondingly, heterogeneities in human mobility patterns can explain more of the variability in regional epidemic data than analyses which posit identically mixed host populations [111, 127, 128] and recent metapopulation models of disease spread have been developed to describe recurrent host movements and retain information on individual identity [129, 112, 130].

An important direction for future work lies in understanding what signals mobility patterns leave in city-level epidemic data when other important factors are integrated into the analysis, such as inter-city variation in age distribution,

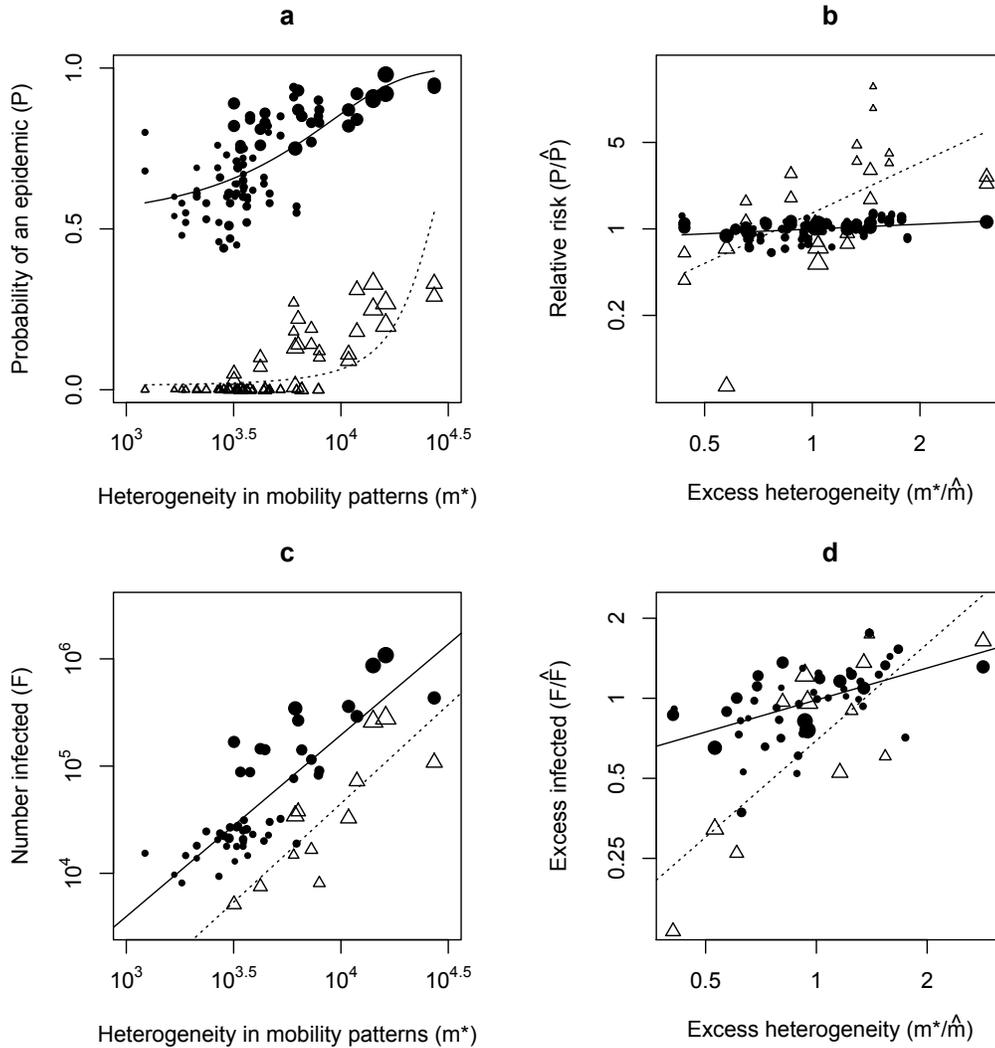


Figure 4.3: **Epidemic dynamics as a function of heterogeneity in human mobility patterns.** **A:** Probability that a single infection will spark an epidemic in 48 cities with different levels of organization in their commuting patterns, calculated from 100 simulations for each city for transmissibilities of $\lambda = 1$ (triangles) and $\lambda = 10$ (circles). Point size is proportional to $\log_{10} N$. Lines show logistic regression controlling for transmissibility. **B:** Relative risk of an epidemic as a function of excess heterogeneity in mobility patterns. The statistical model for \hat{P} is $\text{logit} \hat{P} = x_{\lambda} \log N + w_{\lambda}$. Lines show linear regression controlling for transmissibility. **C:** Final number infected is positively correlated with the level of heterogeneity in mobility patterns. Lines show fits of linear regression on log-transformed variables; $\lambda = 1$ (triangles, dashed line), $\lambda = 10$ (circles, solid line). Point size is proportional to $\log_{10} N$. **D:** This effect persists when the effects of population size on F and m^* are removed.

immune history, or movements within and between cities not accounted for by the commuting data. For instance, we simulated the epidemic dynamics of each city independently, but movement of individuals among cities also affects epidemic dynamics [131]. Another area for improvement is that the commuting data we used represents only the movements of working adults, but the dynamics of many respiratory infections depend on transmission among children [132]. In addition to having different levels of susceptibility, younger age groups can have different mobility and contact patterns [114]. We hypothesize, however, that the transmission model described here approximates (albeit imprecisely) the presence of children by creating contacts among working adults who reside in the same CT. In addition, it is plausible that the mobility patterns of children are similar across cities, so while transmission among children is important, the mobility patterns of workers lead to differences among cities. And differences in influenza dynamics among US states have been partially explained using only the movement patterns of workers [111].

Our results provide empirical support for the potential importance of contact heterogeneity at the intra-city scale, and show new evidence that successfully forecasting epidemics in cities may require us to identify differences in intra-city mobility patterns among cities of similar sizes. Conversely, increasing numbers of infected in larger populations are not necessarily caused exclusively by increases in the number of potential hosts. Instead, increases in the level of heterogeneity in human mobility patterns in larger cities are sufficient in themselves to significantly increase the risk and final size of epidemics. In the face of limited infrastructure for rapidly implementing quarantine and vaccination policies to control the spread of emerging pathogens, an empirical link between human mobility patterns and disease incidence at the scale of individ-

ual cities may allow more effective containment strategies, which exploit predictable inter-city differences in the rate of disease transmission. Analyses that connect detailed information on human contact patterns with city-level disease data are required in order to test the importance to real epidemics of the systematic differences in mobility patterns we have described here [133].

4.5 Acknowledgements

This chapter is based on the article: Dalziel, B.D., Pourbohloul, B., and Ellner, S.P. 2013. Human mobility patterns predict divergent epidemic dynamics among cities. *Proceedings of the Royal Society B: Biological Sciences* 280. BDD, BP, and SPE designed the study. BDD carried out the analysis. BDD, BP, and SPE wrote the paper.

CHAPTER 5
POPULATION DYNAMICS, EVOLUTION, AND CONTROL OF
EMERGING CANINE INFLUENZA VIRUS IN THE UNITED STATES

5.1 Introduction

Respiratory pathogens that emerge as the result of host-range shifts can cause serious epidemics in humans, livestock, and wild animals [134, 135, 136]. Two recent pandemics in humans, Severe Acute Respiratory Syndrome (SARS) in 2003 and H1N1 influenza in 2009, involved host-range shifts in respiratory zoonotic viruses [137, 138]. Importantly, however, such cross-species transmission events do not always result in pandemics. Rather, zoonoses emerging in new host species tend to have patchy prevalence in space and time. As a result, the probability that an emerging zoonosis will take hold in a new host population has been difficult to assess a priori, limiting our capacity to use targeted interventions to avert pandemics before they happen [139].

Several hypotheses explain the patchy distribution of a pathogen after a host-range shift. First, the emerging pathogen may be poorly adapted for replication and onward transmission in the new host population. In this case, the disease will have a lower basic reproductive number (R_0 — the number of secondary infections caused by a typical infected individual in an entirely susceptible population) in the recipient host than its recent ancestor in the donor host. Inefficient transmission following a spillover event may lead to “stuttering chains” of infection marked by patchy patterns of disease prevalence interspersed with stochastic fadeouts. As R_0 declines toward 1.0 it becomes increasingly likely that the new disease will die out altogether; conversely, a higher

R_0 increases the chances that a single spillover event will lead to self-sustaining spread and an epidemic in the new host.

The second (and likely overlapping) hypothesis that explains prevalence heterogeneity in emerging pathogens is that host populations exhibit both demographic and environmental variability. In smaller host populations random variations in the timing of individual birth, death, immigration and emigration events, as well as random variation in the timing of individual infections, can have a profound effect on epidemic dynamics [140, 141, 142]. Emerging pathogens that result from spillover into new hosts are by definition initially confined to a small population, in the sense that the first infected individual(s) will have limited numbers of potential contacts to whom they can spread the disease. This makes the epidemic dynamics of emerging pathogens inherently stochastic [21, 102].

Finally, evolutionary change in emerging pathogens can affect both their basic reproductive number, and their response to demographic and environmental variability. Pathogen evolution can push R_0 upward toward or above 1.0 through repeated spillover events from the reservoir population, or through a chain of transmission in the new host, either of which could lead to the selection of host-adaptive mutations. The occurrence of multiple outbreaks over time may also increase the likelihood that the pathogen evolves toward a point when it can be self-sustaining in the new host [142].

Recent analytical frameworks that unite the ecological and evolutionary dynamics of host-pathogen interactions can help to identify the processes that drive epidemiological and phylogenetic patterns during and after host range shifts [142, 141, 143]. We employ this approach to study the population dynam-

ics, evolution and control of equine-H3N8 derived canine influenza virus (CIV) in the US. CIV began from the transfer of a single H3N8 equine influenza (EIV) to dogs from horses around 1999. Direct descendants of that virus been circulating continuously in dogs since that time [144, 145, 146]. CIV was first recognized as the cause of disease in greyhounds in a training facility in Florida in 2004 and was transferred to various states in the US with the racing greyhounds, eventually spreading to other breeds [144]. The hemagglutinin (HA) sequence in CIV was genetically distinct from EIV by 2004, forming a monophyletic group with a significant difference from its recent ancestor in EIV under pairwise nucleotide sequence comparison [144]. There have been no reports of recombination of the CIV with any other influenza viruses. Notably, there is also no evidence of CIV transfer back to horses, nor onward to humans. Furthermore, although some other H3N8 EIV spillovers from horses into dogs have been reported, those consisted only of single infections or small outbreaks that died out quite quickly [147].

Although CIV can readily transmit among dogs its prevalence remains patchy [148, 149, 150], and it is enzootic in some regions of the US, while the virus has so-far failed to establish after outbreaks outside of those enzootic regions [151]. The overall seroprevalence in the pet dog population appears to be low (3% or less depending on the region), with visits to canine daycare being a risk factor [148, 150]. CIV enzootic regions are typically associated with large animal shelters [152], and the movement of the virus to different parts of the US is most likely associated with the transport of infected shelter dogs to facilities in other regions where they may be more readily adopted.

In contrast to CIV, its recent ancestor the H3N8 EIV has been circulating

widely in horses since before 1963 when it was first reported in Florida, having most likely been introduced with horses from South America [153]. The virus appears to spread continuously in many parts of North and South America, Europe, and Asia [154, 155, 156, 157]. EIV has been introduced into countries that were previously free of the virus, including Australia and South Africa, causing significant outbreaks that extended over large distances, although those were controlled and the virus eradicated [158, 154]. Data from an outbreak in an unvaccinated population of racehorses places R_0 for EIV at 10.18 (95% confidence interval: 9.57 - 10.89) in that context. In contrast, the reproductive rate of EIV in vaccinated populations of racehorses has been estimated to be between 1.4 and 2.3 [159]. EIV has experienced marked evolution in all gene segments since it emerged, with evidence of antigenic variation in the HA gene, including geographic patterns in antigenic variation [160, 154, 161].

Although CIV and EIV are closely related, their epidemiology and evolutionary dynamics differ, with EIV seemingly more successful, and less heterogeneously distributed. Moreover, EIV continues to spread despite considerable control measures (particularly vaccination) whereas CIV retains a patchy distribution in the absence of significant control measures. Studying the ecology and phylogeny of CIV since its recent emergence from EIV will therefore help to elucidate how host demography, disease dynamics, and pathogen evolution combine to determine the prevalence patterns and risk posed by emerging zoonotic pathogens.

Here we use standard models of pathogen spread and diversification to combine individual-level data on the intake, output, and transfer rates of dogs among US animal shelters, with CIV gene sequence data and available

seroprevalence estimates, to examine what processes control transmission and prevalence patterns in CIV, and to explore possible strategies for its eradication. We hypothesize that CIV has a lower R_0 than EIV, but persists through the presence of transmission hotspots, which rescue stuttering chains of transmission that fade out in other populations. The putative hotspots are large animal shelters in major metropolitan areas. After estimating R_0 from all available data we ask: are the population sizes of small shelters small enough to make fadeout likely, and significantly more likely than in large shelters? Conversely, do large shelters have good prospects of maintaining CIV in an enzootic state?

5.2 Methods

5.2.1 Model

Our analysis is based on an *SIR* framework that models changes over time in the number of dogs in a shelter who are susceptible (S), infected (I), or removed (recovered and thus immune; R). Below we expand the model to consider multiple shelters linked through the transfer of dogs, and to incorporate the dynamics of an intervention program with a live-attenuated vaccine administered to dogs on arrival.

We assume dogs arrive at a shelter of a given size at a rate of μ dogs per day. Dogs leave at a per-capita rate of δ per dog per day, regardless of their state, so the mean residence time in a shelter is $1/\delta$ days. The number of dogs in a shelter, $N = S + I + R$, is equal to μ/δ at equilibrium. Arrival and departure rates are estimated empirically using individual-level records from 13 animal

shelters of varying size across the US. The records comprise a total of 124,519 dogs, recording the date each individual arrived and left the shelter. In 8 of the 13 shelters, the data included whether or not the departure of the dog represented a transfer to another shelter. Arrival rate, μ , for a shelter was estimated as the median number of dogs arriving in that shelter per day. Departure rate, δ , for a shelter was estimated as the inverse of the median length of stay of dogs in that shelter. When estimating arrival and departure rates we excluded dogs who were admitted to the shelter in response to a euthanasia request, as these dogs had systematically shorter residence times. We also excluded dogs whose length of stay was greater than 40 days, as these represented rare atypical cases (see Figure 5.2C,D).

We assume that dogs in a shelter have a constant rate of contact per day with other dogs where the contact would be capable of spreading infection if one of the dogs were infected. Assuming that contact between any pair of dogs in the shelter is equally likely, the force of infection is given by $\lambda = \beta P$, where β is the contact rate and $P = I/N$ is the current prevalence of CIV in the shelter [162]. The rate of appearance of new infections is given by λS , and susceptible dogs contract the disease an average of $1/\lambda$ days after entering the shelter. In this framework, the basic reproductive number of the disease is $R_0 = \beta/(\gamma + \delta)$, and the disease only persists in the long run if $R_0 > 1$, in which case equilibrium prevalence is given by

$$P = \frac{\delta}{\gamma + \delta} \left(1 - \frac{1}{R_0} \right) \quad (5.1)$$

which is bounded above by $\delta/(\gamma + \delta)$ as R_0 becomes large (see Appendix C).

The infected class in our model represents the number of dogs with non-zero viral loads, rather than those exhibiting clinical symptoms. Thus, we avoid in-

cluding latent or asymptomatic classes in our model. We set $\gamma = 1/7$ because viral shedding continues for approximately seven days after inoculation [144]. Seroreconversion for dogs infected with CIV also happens at approximately 7 days [144]. Equilibrium seroprevalence is then given by R/N .

Variation among individuals in time of infection, recovery, arrival, and departure causes variations in disease prevalence around the predicted long-term average. These excursions from mean prevalence carry with them the risk of visiting zero prevalence, leading to stochastic extinction of the disease. This demographic stochasticity becomes increasingly pronounced in smaller populations. However, the critical population size below which disease dynamics begin to significantly diverge from the long-term average through stochastic fadeouts depends upon R_0 , and upon the turnover rate in the population. We parameterize the stochastic SIR model with the demographic data to test the impact of demographic stochasticity on the spread and maintenance of CIV in animal shelters. We implement the model in continuous time at the level of individual dogs using the Gillespie algorithm [163].

We estimated a posterior distribution for R_0 given point seroprevalence data and demographic data by using a Markov Chain Monte Carlo (MCMC) method, as follows. From the stochastic SIR model we simulated point seroprevalence samples by observing the seropositivity of n randomly selected dogs from the population at a given time. Point seroprevalence has the property of being normally-distributed about the long-term equilibrium value given by the mean-field model in our simulations (see Figure 5.3B). We then seek the posterior distribution of an unknown equilibrium seroprevalence at an actual shelter, given a point seroprevalence estimate there. We estimate this distribution by sampling

from the Gaussian distribution of deviations between point seroprevalence estimates and equilibrium seroprevalence, using the Metropolis-Hastings algorithm [164]. Convergence was easily achieved using 10 chains run for 100000 steps each, with a burnin of 10%, and keeping every 100th step. From the posterior distribution of equilibrium seroprevalence, we map to a posterior distribution for R_0 by inverting Equation 5.1.

5.2.2 Vaccination - inactivated or modified live intranasal

The model with vaccination dynamics includes two more compartments, counting the number of dogs in each shelter who are vaccinated (V), and the number of dogs who are infected despite vaccination (W). Vaccination reduces a dog's susceptibility to infection by decreasing the probability that a virus population initially transferred through infectious contact will enter a phase of exponential growth, prerequisite to significant viral shedding and clinical symptoms [165]. By reducing viral load and viral shedding, vaccination reduces the risk of infection in vaccinated dogs and reduces the infectiousness of a dog who becomes infected despite vaccination. Vaccinated dogs thus experience a reduced force of infection $\varepsilon\lambda$, $0 \leq \varepsilon \leq 1$, and, if they become infected, contribute to the force of infection at a reduced rate $0 \leq \omega \leq 1$, leading to an overall force of infection of $\lambda = \beta(I + \omega W)/N$ in population which has W vaccinated individuals who have nonetheless become infected.

Dogs transition from S to V at a rate of α per dog per day. The term $1/\alpha$ measures the average time after entry/vaccination that a dog experiences the vaccine-associated decrease in risk of infection from other dogs, and decreased

infectiousness if they do become infected. Vaccination changes mean dynamics by reducing R_0 by a factor of $1 - K$, where K is effective vaccination coverage. K is given by $(1 - \kappa)V/N$, where $\kappa = \varepsilon\omega$ expresses the failure rate of the vaccine, ranging from 0 for perfect vaccine, to 1 for an entirely ineffective one (see Appendix C). We use a step function for κ as a function of α , where κ goes from 1 to its post-vaccine value at $1/\alpha$ days.

We also model the effects of a control strategy equivalent to inoculating some dogs with a perfect vaccine, or to quarantine that partially or completely stops the flow of susceptible dogs into the shelter. We do this by replacing susceptible dogs with removed ones in the intake stream. Reducing the proportion of susceptible dogs in the intake stream to $0 \leq \theta \leq 1$, while $1 - \theta$ are already removed, has the same effect as reducing R_0 to θR_0 .

5.2.3 Metapopulation dynamics

The metapopulation model expands the stochastic SIR model for a single shelter to describe multiple shelters whose dynamics are linked by the transfer of dogs. As above, the model is implemented at the level individual dogs using the Gillespie algorithm. Thus at each point in continuous time, each individual in the model has a disease state (S, I, R, V , or W) and a location in a given shelter. The metapopulation is composed of shelters that vary in dog population size, intake rate and output rate by sampling with replacement from the demographic data. Transfer probabilities are also based on the demographic data (see Appendix C).

Although the CIV phylogenies show geographic localization (see Figure 1),

the metapopulation model is spatially implicit, consistent with the level of detail in the demographic data we used. That is, we currently do not have enough data to parameterize even a simple spatially explicit model of dog movement amongst shelters (such as a gravity model). However, even without including spatial structure in transfer patterns, the metapopulation model reproduces hotspot dynamics, based on transfer hierarchies driven by differences in shelter size alone (see Figure 6). That is, large shelters receive the infection earlier, and maintain it for longer, leading to a predicted spatially patchy distribution in CIV prevalence, consistent with geographic localization. Thus we hypothesize that each geographically distinct clade in the phylogeny is associated with one or more large animal shelters.

5.2.4 Phylogenetic analysis, estimates of R, and phylogeography

We obtained all available CIV HA1, NP and M gene segment sequences from GenBank and by sequencing samples provided by the Animal Health Diagnostic Center (AHDC) at Cornell University. For the sequencing of the virus samples obtained from AHDC we extracted viral RNA using Qiagen viral RNA mini kit and synthesized cDNA using Avian Myeloblastosis Virus (AMV) reverse transcriptase and influenza universal primer Uni12. Three gene segments, HA1, NP and M, were then amplified by PCR with gene specific primers (primer sequences are available upon request) for all samples. The PCR products were purified using EZNA Cycle-Pure Kit and sequenced by the Sanger method.

All sequences were aligned by MUSCLE v3.8.31[166] using default parame-

ters, followed by manual adjustment. Phylogenetic trees of each gene were then estimated using the maximum likelihood (ML) available in PhyML 3.0 [167] and assuming the general time-reversible reversible (GTR) model of nucleotide substitution and a gamma distribution of among-site rate variation with 4 rate categories (i.e. the GTR+I+ Γ 4 model of nucleotide substitution) with SPR branch-swapping. The robustness of the phylogeny was estimated using 1,000 bootstrap replicates. Because of their greater availability, the analyses of evolutionary dynamics and phylogeography were only performed on the HA1 gene (see below).

We used the HA1 sequences to estimate the effective reproductive number of the virus, denoted R . By a common abuse of notation, R refers both to the effective reproductive number of the virus, and to the number of removed individuals in the population under an SIR model. However it will be clear from the context which is being referred to. To estimate the effective reproductive number R from the CIV sequence data we utilized a total of 94 HA1 sequences (alignment length = 1032 nt) sampled from various locations (states) in the USA (Colorado, New York, Pennsylvania, Florida, California, Kentucky, Wyoming, Philadelphia, South Carolina, Virginia, Vermont, Connecticut, Texas, and Iowa) between 2003-2013. This data set included 40 sequences sampled from dog shelters in New York between 2005-2012, which were analyzed separately using the same protocols. First, we estimated the mean (and credible intervals) of R in both data sets using the epidemiological birth-death method [168] available in BEAST v1.7.5 [169]. This analysis utilized the Hasegawa-Kishino-Yano (HKY) model of nucleotide substitution and a gamma distribution of among-site rate variation (HKY+ Γ 4). To account for any rate variation in the data an uncorrelated lognormal relaxed molecular clock model was employed. Using the

Bayesian Markov Chain Monte Carlo (MCMC) framework available in BEAST, 100 million steps were run, sampling every 10,000 and removing 10% as burn-in. Second, temporal changes in R were estimated using the more complex serial-sampled birth-death (SSBD) model [143], available in BEAST v2.0 [170], again using the HKY+ Γ 4 model but this time (to ensure statistical convergence) employing a strict molecular clock with a uniform distributed clock rate of 2×10^{-3} (1×10^{-3} - 3×10^{-3}) nucleotide substitutions per site, as this was found to be best-fit to the data in epidemiological birth-death method. The MCMC was again run for 100 million steps, sampling in the same way as described above.

To determine whether CIV was more clustered on the phylogenetic tree by US state of sampling than expected by chance alone, we employed the Association Index (AI), Parsimony Score (PS) and Maximum Clade size (MC) phylogeny-trait association statistics incorporated within the Bayesian Tip-association Significance testing (BaTS) program [171]. Traits were defined as the US state of sampling each sequence and phylogenetic uncertainty in the data was incorporated by basing estimates on the posterior distribution of trees obtained from the BEAST analysis (epidemiological birth-death method) described above. In all cases, 1000 random permutations of sampling locations were undertaken to create a null distribution for each statistic.

5.3 Results

5.3.1 Phylogenetic Structure of CIV in the USA

To put the CIV sampled from animal shelters in a wider geographical context, and to reveal movement of the virus on a continental scale, we determined the HA1, M and NP gene sequences of recent CIV isolates, and combined those with sequences available on GenBank through phylogenetic analysis. Viruses were from various US states including Colorado, New York, Pennsylvania, Florida, California, Kentucky, Wyoming, Philadelphia, South Carolina, Virginia, Vermont, Connecticut, Texas, and Iowa. The most striking result of this analysis is that CIV exhibits a marked geographical clustering by US state of sampling. In particular, distinct clades were observed in New York, Pennsylvania, and Colorado, which represent our largest sampling sets (Figure 5.1). In addition, many of the viruses from the North East states of Vermont, Connecticut, and New Hampshire clustered with the viruses from New York, as well as a virus from South Carolina, suggesting that those viruses were derived from New York. This geographical clustering was confirmed in AI and PS phylogeny-trait association statistics [171], with significantly more clustering by US state of origin than expected by chance alone across the data set as a whole ($p = 0$). In addition, the MC statistic reveals significant ($p < 0.001$) clustering in the individual states of Colorado, New York, Pennsylvania, Vermont and Wyoming.

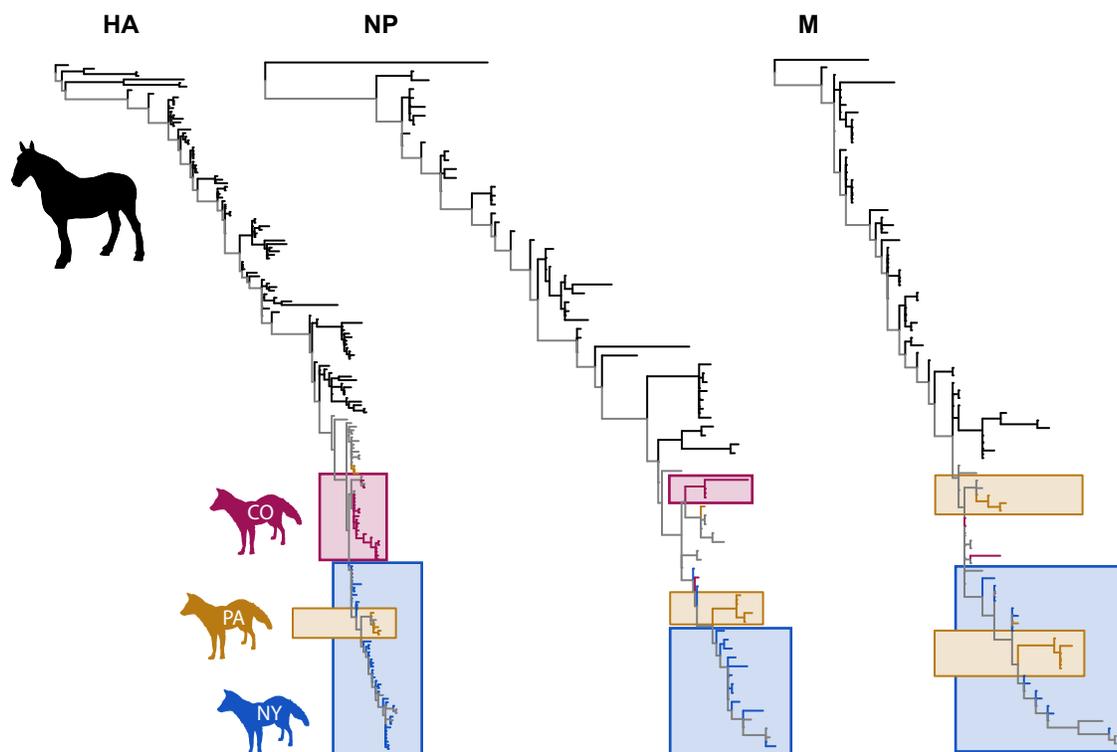


Figure 5.1: **Phylogenetic trees of HA1, NP and M sequences for EIV (black) and CIV (colors).** Boxes surround CIV clades comprising two or more samples from the same US state. Branches leading to CIV samples from the same location are colored by location (New York, blue; Pennsylvania, orange; Colorado, red). Branches leading to CIV samples from multiple locations are colored grey.

5.3.2 Epidemiological Dynamics of CIV in Shelter and Domestic Dogs

Next, we investigated the epidemiological dynamics of CIV at the local scale, in animal shelters. The majority of animal shelters in the US house relatively small populations of dogs—the median dog population size in our sample of shelters is 43—but a few shelters are much larger, housing hundreds of dogs. In precise terms, the distribution of dog population sizes in our data is close to a negative binomial distribution with mean 71.23 and standard deviation 82.24 (Figure 5.2A), which indicates significant overdispersion in population sizes relative to a homogeneous Poisson model. This overdispersion in host population size is a potentially important characteristic for the epidemiology of CIV because it indicates the presence of a few extraordinarily large shelters where a pathogen might persist more easily than in a host population of average size. Larger shelters are fueled primarily by higher intake rates (Figure 5.2B), as median residence time of dogs does not vary significantly among shelters of different sizes (Figure 5.2C). The median residence time of dogs across all shelters is gamma distributed with a mean of 9.88 days and a standard deviation of 8.22 days (Figure 2D). Transfer rates among shelters appear relatively low—among the eight shelters in our demographic data for which there was transfer information the median proportion of outcomes that were transfers is 0.067 and the mean is 0.1. Transfer probability is not correlated with dog population size in our data.

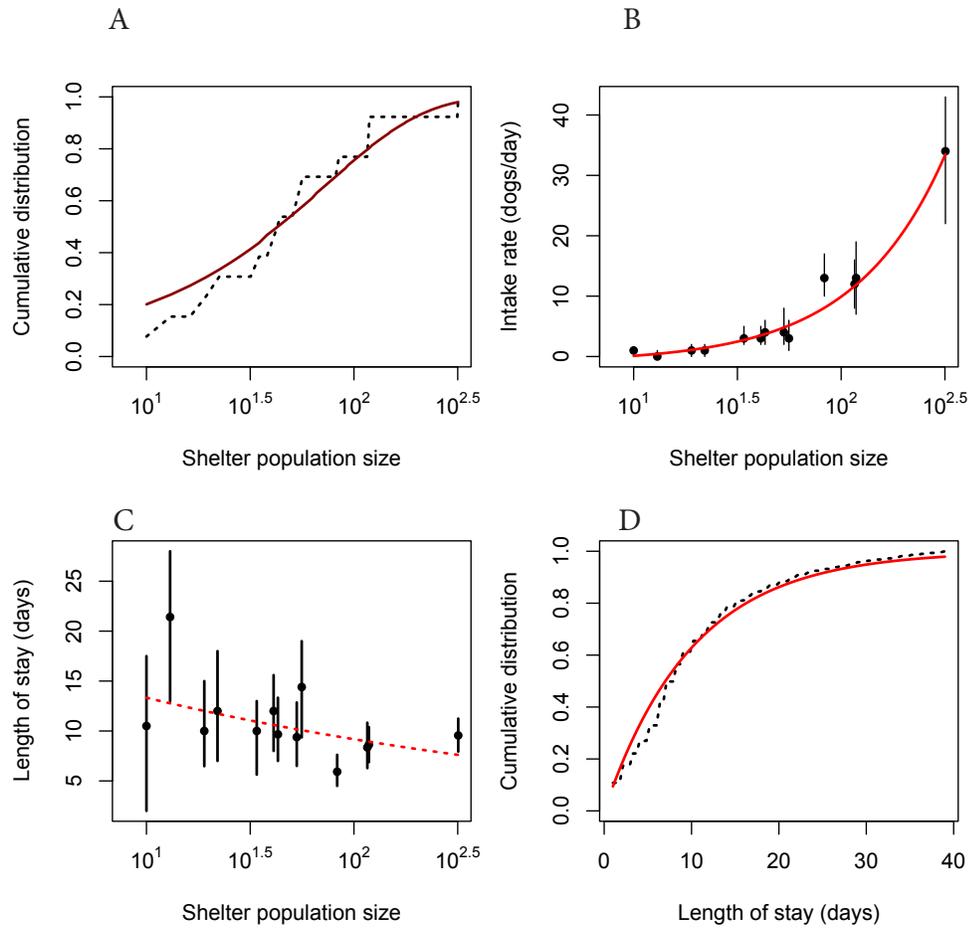


Figure 5.2: **Demography of dogs in US animal shelters.** **A:** Cumulative distribution of median population size in each shelter (dashed line) compared to a negative binomial distribution fitted to the data (solid line). **B:** Intake rate as a function of population size. Points show the median value for each shelter and vertical lines enclose the interquartile range. Line shows fit by linear regression to log-transformed median intake rates. **C** Length of stay as a function of shelter size. The slope of the dashed line does not differ significantly from 0. **D** Cumulative distribution of length of stay across all shelters (bars) compared to an exponential distribution with mean rate $1/9.88 \text{ days}^{-1}$ (solid line).

Most dogs arriving to shelters are susceptible to CIV [148, 152]. The arrival rate of susceptible dogs places an upper limit on CIV prevalence by continual dilution with uninfected individuals (Figure 5.3A). At the same time, the empirical data on arrival and departure rates indicate that large high throughput facilities could fuel persistent infections by guaranteeing a supply of new susceptible individuals, whereas in smaller facilities random variation in arrival, departure, residence times, infection and recovery times are likely to cause stochastic fade-outs of the disease. The magnitude of the impact of demographic stochasticity depends on R_0 and on the demography of the host population. For CIV in animal shelters the stochastic simulations parameterized with demographic data reveal that the impact of demographic stochasticity is considerable; the majority of shelters are too small to maintain the virus in the long term at its present rate of transmission (Figure 5.3B and Figure 5.4A).

A point seroprevalence estimate of 0.41 [152] from a large shelter where CIV is enzootic, combined with the demographic data on dog intake and outcome rates, yield a mean estimate for R_0 of 3.9 for CIV in large animal shelters. The posterior distribution of R_0 has a median of 3.3, and a highest probability density (HPD - the central 95% of the posterior distribution) interval of extending from 2.0 to 8.9 (Figure 5.3C). Our estimates for the effective reproductive number R , for both the USA as a whole, and New York specifically, estimated using a phylodynamic method on HA1 sequence data, show considerable temporal variation (Figure 5.3D). At the time when CIV was first recognized in 2004 the posterior distribution of the effective reproductive number R (the average number of secondary cases actually produced by an infectious individual at a given time during the epidemic) roughly matches that of R_0 , indicating an exponential spread rate of the disease. During the period 2004-2008 R drops to a value of

1.0. A similar demographic signature—declining R to a value close to 1—was observed in the New York data set. However, the wide 95% HPD values, reflecting the relatively small sample size (and with non-random sampling across the US), means that caution should be exercised when interpreting the temporal trend in effective reproductive rate. Across the USA as a whole the mean estimate of R is currently 1.02 (95% HPD = 0.79,1.26), with a similar figure found in New York ($R = 1.06$, 95% HPD = 0.72, 1.47). The low R observed toward the present suggests that CIV has now reach an equilibrium, where stochastic fade-outs associated with outbreaks are balanced with new infections in the large animal shelters where it is enzootic.

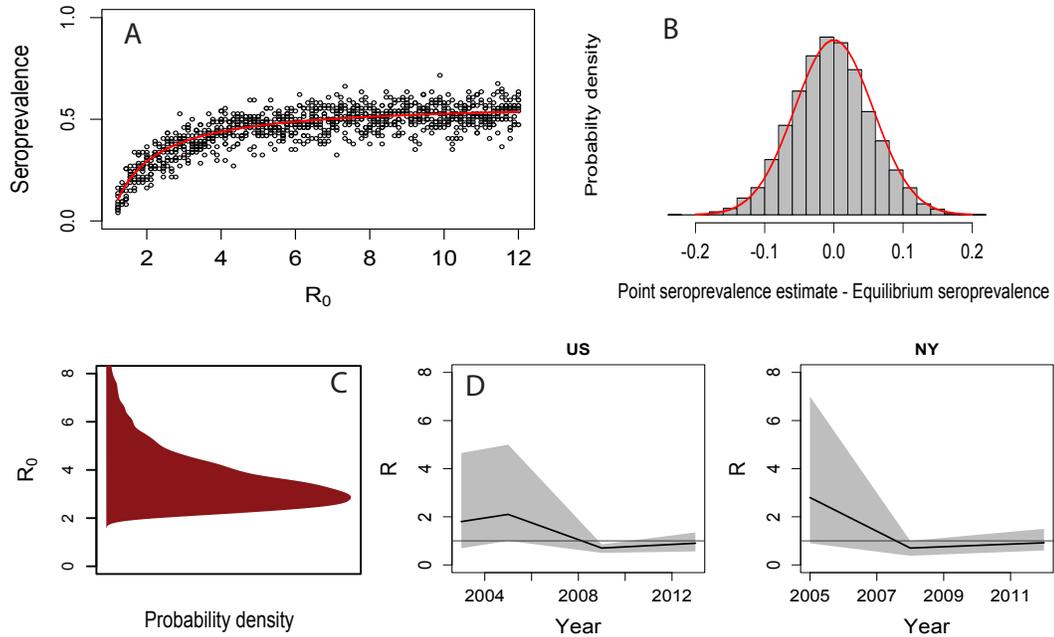


Figure 5.3: **Seroprevalence, R_0 and R for CIV, estimated from host demographic data, seroprevalence data, and molecular data.** **A:** Saturating relationship between seroprevalence and R_0 in a stochastic SIR framework, parameterized from the shelter intake and output data. Red line shows equilibrium seroprevalence predicted by the mean-field model. Points show point seroprevalence estimates from the stochastic simulations, where 74 dogs are sampled at random in a shelter with an average dog population of 134, corresponding to [152]. **B:** Deviations of point seroprevalence estimates from the long-term average (bars) compared to a normal distribution (line). **C:** Posterior distribution of R_0 based on an observed seroprevalence of 0.41 in [152]. **D:** R for CIV, estimated by fitting a birth-death skyline phylodynamic model to HA1 gene sequences. The black line shows the mean estimate while the grey shaded shows the highest probability density (HPD) range, encompassing 95% of the credible set of sampled values.

5.3.3 Populations that sustain viral transmission

Using the shelter demography data, we simulated CIV outbreaks in shelters of a realistic range of sizes, intake rates and output rates, and for varying levels of R_0 . From these simulations we estimated the probability that a shelter (of a given size) infected with a CIV virus (of a given R_0) could maintain the virus for 100 days. The response surface for this experiment yielded a cutoff curve in the N - R_0 plane, below which fadeout was almost certain and above which persistence was almost certain (Figure 5.4A). Interestingly, the posterior distribution for R_0 and N straddles the boarder between persistence and stochastic fadeout. The demographic and seroprevalence data thus indicate that CIV cannot persist in the majority of shelters (the median N/R_0 combination is below the cutoff for persistence) but can persist in certain larger shelters (the joint distribution of N and R_0 extends beyond the cutoff). Figure 5.4B shows the effect of reducing the inflow of susceptible dogs to effectively reduce R_0 from 3.9 to one through a vaccination or quarantine program.

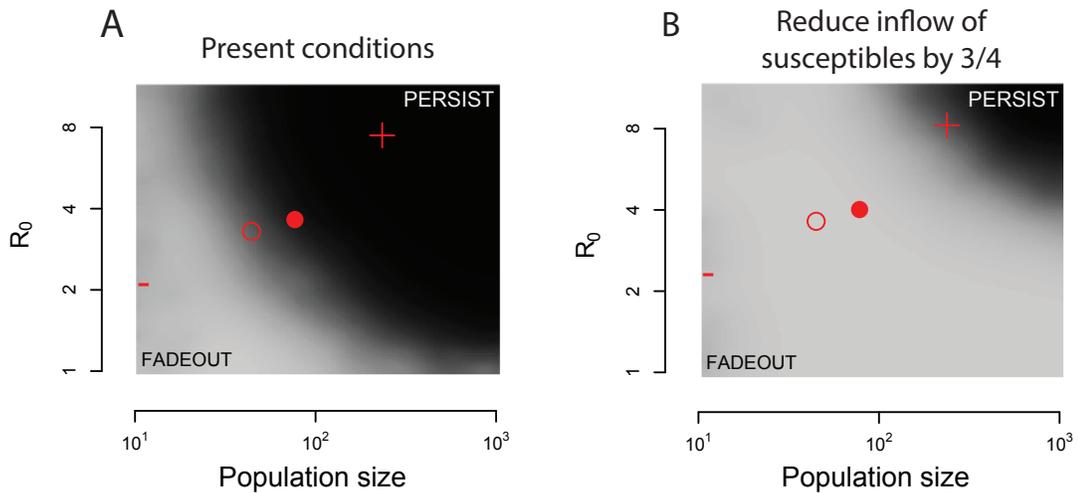


Figure 5.4: **Demographics, persistence, spread rate and possible eradication of CIV.** **A:** Dog population sizes in animal shelters and within-shelter spread rates at which CIV can persist for at least 100 days according to present intake and output rates. The surface shows a smoothed version of the outcome of 1000 simulations conducted at random points within the plane described by the figure. Darker shades correspond to higher probabilities of persistence. Red symbols show features of the empirical joint distribution for dog population size and R_0 in shelters (see Figures 5.1 and 5.2), including the median (hollow circle), mean (filled circle), 2.5th percentile (minus sign) and 97.5th percentile (plus sign). **B:** Results of an intervention that reduces the arrival rate of susceptible individuals at a shelter to $1/3.9$ its current value, equivalent to reducing the mean estimate for R_0 to 1.

5.3.4 Control and Eradication Strategies

According to our analysis (based on the stochastic *SIR* model parameterized with the demographic data and estimates of the basic reproductive number from observed seroprevalence), a vaccination program with a live attenuated influenza vaccine (LAIV) could eradicate CIV within 1-2 months if the vaccine is administered to dogs immediately upon arrival to the shelter, and removes them from the chain of transmission within 24 hours with 85% probability (hereafter “vaccine efficacy”; Figure 5.5A). A vaccine with an efficacy of 75% might also efficiently eradicate CIV from isolated shelters, but transfers of dogs between shelters at the observed mean rate will allow CIV to persist through connected chains of outbreaks (Figure 5.5B). Vaccine efficacies of 65% or less reduce prevalence but are not predicted to lead to CIV eradication (Figure 5.5C).

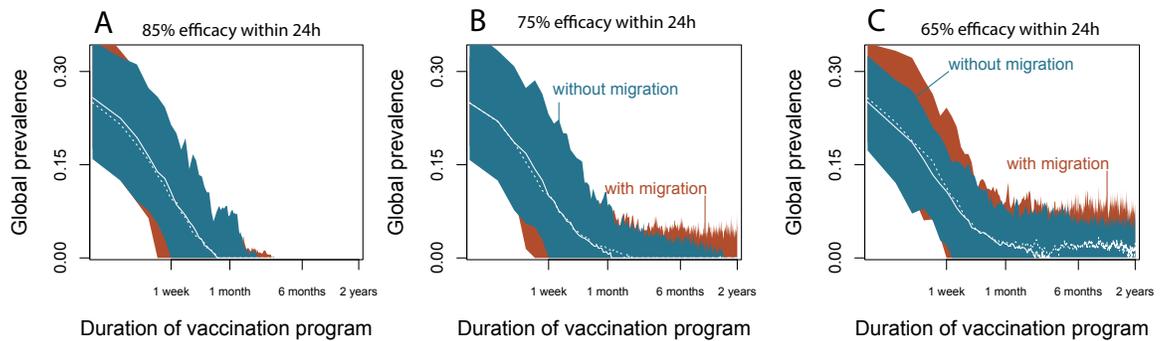


Figure 5.5: **Predicted performance of a control program using a live-attenuated vaccine administered to dogs on arrival in US animal shelters.** **A:** A vaccine that removes individuals from the chain of transmission with 85% probability ($\kappa=0.15$) within 24h ($\alpha=1$ day) is predicted to eradicate CIV from shelters within six months. The simulations used 100 shelters with dog population size, intake rate, and outtake rate jointly sampled with replacement from the shelter demographics data, and $R_0 = 3.9$. White lines show medians and shaded areas enclose the 5th to the 95th percentiles of the simulation data. **B:** Decreasing vaccine efficacy to 75% can still achieve eradication in isolated shelters (blue region, solid line), however shelters that transfer dogs amongst themselves at the observed mean rate of $\tau = 0.1$ would preserve CIV in a few shelters despite the vaccination program (red region, dashed line). **C:** Further decreases in vaccine efficacy make eradication significantly less likely, particularly if shelters are connected through the transfer of dogs.

Turnover rates in most shelters are too high for an inactivated vaccine to be effective, because inactivated influenza vaccines typically take more than a week to generate immunity. Since the expected residence time of a dog in an animal shelter is around 10 days, most dogs would leave the shelter, and have been part of the chain of transmission, by the time an inactivated vaccine took effect. Other control measures (e.g. quarantine, decrease in population size or changes in population structure, anti-viral drugs) or combinations that accomplished the same level of infection decrease would also have qualitatively similar effects to vaccination.

We also used our epidemic model to explore the passage of CIV from an infection in one large shelter to other shelters through the transfer of dogs (Figure 5.6A). The hotspot dynamics predicted by our model show regularities in the way CIV spreads outward from a single shelter. Large shelters are predicted to receive the infection earlier, as well as maintaining it for longer, creating a wave in the population size—time-of infection plane (Figure 5.6B). The probability that a single infection introduced to a susceptible shelter would start an epidemic that persisted for at least 100 days increases with population size (Figure 5.6C). For the median population size of 43 dogs the probability was approximately 0.5.

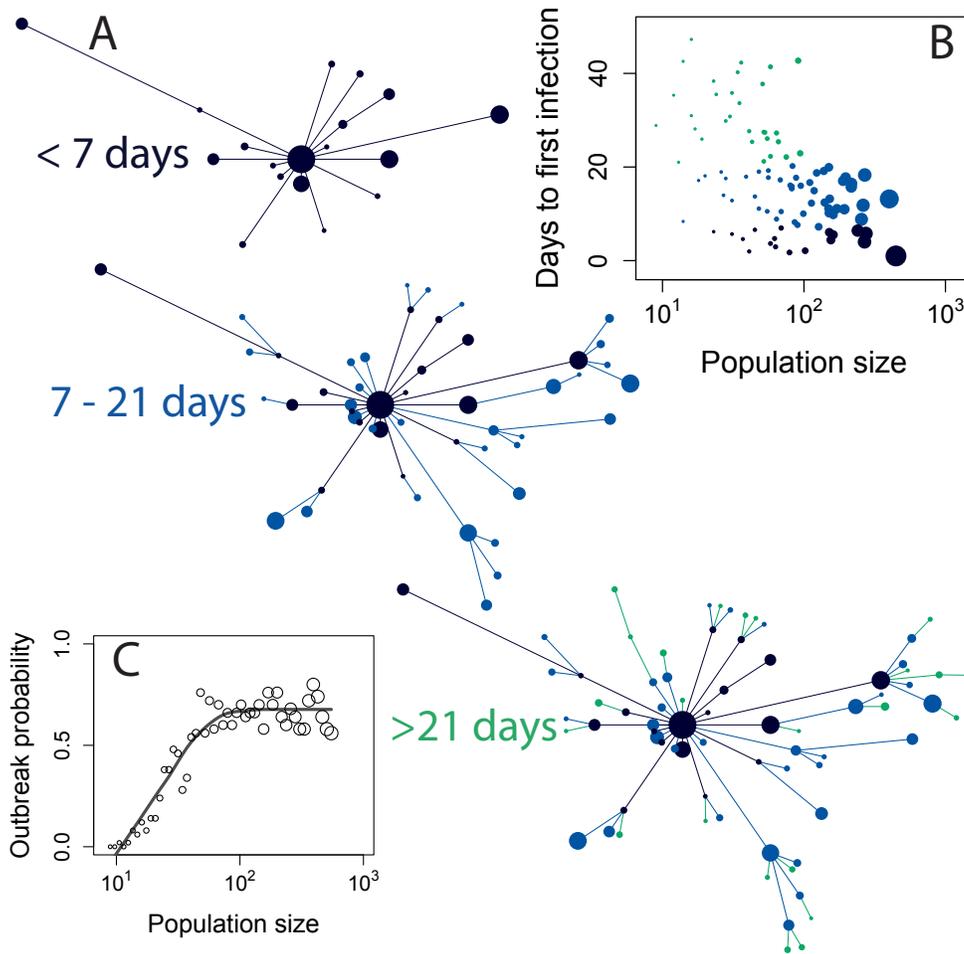


Figure 5.6: **A CIV invasion over multiple shelters, starting with an infection in a single large shelter.** **A:** Each vertex represents an animal shelter with dog population size proportional to the area of the circle. Edges show transfer of infection from shelter to shelter over time through the movement of infected dogs. Edge lengths are arbitrary. The data for this figure were produced by simulating the metapopulation stochastic *SIR* model with 100 shelters for 100 days, starting with a single infection in the largest shelter. Population sizes were sampled with replacement from the shelter data. $R_0 = 3.9$. Transfer probability is set to the mean observed value of $\tau = 0.1$. **B:** Large shelters tend to receive the infection earlier (and more often) following an outbreak at another shelter. **C:** Probability that CIV will persist for 100 days in a shelter of a given size following the introduction of a single infected individual to an otherwise susceptible population.

5.4 Discussion

Since its emergence more than a decade ago, equine H3N8-derived CIV has maintained a patchy distribution, confined to the US and often occurring in sporadic short-lived outbreaks [151]. In contrast, strains of EIV H3N8 have been widespread, with the virus spreading rapidly around much of the world in horses since it emerged around 1963, repeatedly demonstrating its capacity to transmit efficiently among horses, sometimes despite vaccination programs [172, 159]. Our study investigates the processes that underlie heterogeneity in CIV transmission and prevalence to understand the factors that determine whether a zoonotic pathogen will take hold in a new host population following a host range shift, and to examine possible eradication programs that could eliminate the virus from dogs.

Our mean estimate of $R_0 = 3.9$ in the large animals shelters is lower than some estimates for EIV during outbreaks [173], but close to the upper bound for estimates of human influenza transmission [174, 175]. It is also considerably higher than that of pandemic H1N1 influenza in humans in 2009 ($R_0 = 1.4 - 1.6$) which spread within weeks of its first recognition in humans [176]. Variation in R_0 among different viral strains and host species can be difficult to interpret because of the many factors that can affect transmission and removal rates in different settings. However, these comparisons do indicate that CIV has the biological capacity to spread relatively efficiently among dogs given the right conditions in the host population, although these may not exist outside of animal shelters.

Variability in our estimate of R_0 is driven by several factors, including vari-

ation in population size within and among shelters, variation in the residence time of dogs in shelters, and variation in seroprevalence within a shelter over time, due to stochastic fluctuations in the transmission and removal processes. While the scarcity of seroprevalence estimates adds uncertainty to associated estimates of R_0 , the extant data would be difficult to explain with values of R_0 lower than our estimates. This is due to the rapid rate at which infected individuals are replaced by new arriving susceptibles in the high throughput shelters where CIV is enzootic. The low residence time of dogs in large, high-throughput shelters thus indicates (consistent with previous results [152]) that individuals in shelters where CIV is enzootic must acquire the infection within a few days of arriving. This places a lower bound on probable values for R_0 by constraining estimates of the generation time of the infection, at least in the context of large shelters [175].

Most US animal shelters do not provide conditions that favor persistence of CIV in the long term. Rather, the demographic data indicate that most shelters are too small, and import susceptible individuals too slowly to protect CIV from stochastic extinction. Furthermore, observed transfer rates suggest the majority of intakes and outputs are not associated with other shelters, so that shelter-to-shelter transfer has not created an effectively larger population of multiple shelters. While there are many millions of household dogs in the USA, it is likely that those do not have infectious contact patterns that are sufficient to maintain the virus in continuous transmission. Together, these results suggest that the patchy distribution of CIV can be explained primarily by demographic stochasticity in relatively small and/or disconnected host populations, rather than by maladaptation of CIV to transmission in dogs.

Phylogenetic analysis independently supports several key predictions of the forgoing analysis. First, the phylogenetic analyses concur that CIV remains confined to endemic hotspots, with transfers to other regions causing outbreaks that are generally short-lived (and thus fail to establish new branches in the phylogeny outside of endemic locations). This is consistent with the demographic and epidemic analysis showing CIV can only persist in relatively few large shelters (see Figure 5.4A). Second, the mean and HPD interval for R in 2004 (when CIV was first detected) from the phylogenetic analysis roughly matches the distribution of R_0 from the stochastic SIR model and demographic data. The initial phase of an epidemic is usually associated with exponential growth (corresponding to $R = R_0$), and R must always be less than or equal to R_0 by definition. This makes the phylogenetic estimate of R in 2004 a conservative independent assessment of R_0 .

The third concurrence between the demographic and phylogenetic analysis involves the current estimate of $R \approx 1$. While R_0 in our analysis measures the reproductive potential of the disease where it is enzootic, the phylodynamic estimates of R reflects the net spread rate of CIV across the US as a whole, including multiple shelters and the pet population. As such, $R \approx 1$ indicates that on balance CIV is currently failing to persist where it is not already enzootic, which is consistent with both epidemiological observations [151] and the predictions of the demographic analysis which showed that most shelters are too small to support CIV in an endemic state (see Figure 5.4). Finally, the relatively low transfer rate observed among shelters (on average about 10% of outcomes are transfers), combined with the high risk of extinction for a CIV infection when transferred to a new shelter (an average sized shelter would not allow CIV to persist), suggests geographic segregation, which is supported by the phylogenetic analysis

showing distinct viral clades in New York, Pennsylvania and Colorado, for each gene analyzed.

The marked differences in R_0 estimated within dog shelters (> 3) and the current value of R across the USA as whole (≈ 1) also point to different selection pressures in these different ecological circumstances. Specifically, in a high-turnover population like a dog shelter natural selection (for transmission) may favor a high viral titer in a short duration of infection, whereas in nature, because dog populations are far more sparse, selection is more likely to favor a longer duration of infection as this will maximize the chance of dog-to-dog transmission. This suggests that the CIV and EIV viruses may show different replication/transmission trade-offs due to their different host population structures, with very different outcomes for the spread of the viruses, despite their close genetic similarity.

Understanding the conditions under which an emerging enzootic pathogen can maintain itself in a new host population can lead to targeted strategies for control and possibly eradication. Our simulations of CIV control—here we suggest using a LAIV to compete for infection of incoming animals with circulating wildtype viruses—indicate that eradication may be possible, but the success of such an endeavor may depend on the rates at which dogs are transferred between animal shelters. For LAIV with moderate efficacy ($\approx 75\%$), CIV may be eradicated from most shelters, but persist overall through connected chain of outbreaks, with large shelters serving as focal points for staging new infections elsewhere. Practically this suggests that LAIV vaccination programs for CIV will be most successful if implemented at scale across multiple shelters and that participating shelters should maintain vaccination even after CIV has been

eliminated locally.

Our analyses necessarily involve significant uncertainties that would be reduced by more information on CIV prevalence and additional CIV sequence data. A key area of uncertainty is transmission rates between pet dogs within and among households whose contact patterns would differ significantly from those of dogs in animal shelters. Working from first principles, if a proportion p of contacts between infectious and susceptible dogs result in the infection being transmitted, then the probability that an infected individual in the pet population will first transmit the infection on their k th contact with a susceptible dog is $(1 - p)p^{k-1}$, which gives an expected value for k of $1 + p/(1 - p)$. Studies of CIV transmission in comingling trials estimate $p = 0.75$ [149]. This suggests that for a CIV lineage to avoid extinction in pet dogs, the average infected dog must contact $k = 1.33$ susceptible individuals during the time they are infected. Another approach that yields the same result is to recognize that if a proportion p of contacts produce secondary infections, then $R_0 > 1$ requires $k > 1/p$, again translating to approximately two contacts per week for $p = 0.75$. It is evident from common experience that while some pet dogs in the US are highly sociable, others do not frequently interact with other dogs. Therefore, some pet dogs probably achieve the minimum contact rate required to sustain CIV, but other dogs probably do not. Higher contact rates than this minimum would be necessary for CIV to actively spread among pet dogs, and to protect the virus from stochastic extinction in the general dog population.

Despite limitations in the available data, our analyses reveal a coherent view of the ecological and evolutionary dynamics of CIV. After approximately 15 years of continuous circulation among dogs in the US, CIV can be maintained

only in relatively dense host populations with high inputs of susceptible individuals, despite a relatively high reproductive potential in that context. These hotspots are weakly connected by migration, leading to geographic signatures in the CIV phylogenies. Our analysis further indicates that most dog populations are too small or diffuse to independently support CIV at its current level of transmissibility, explaining its current modest reproductive rate ($R \approx 1$), and consistent with the epidemiology of CIV, which is characterized primarily by sporadic short-lived outbreaks outside of enzootic centers [151]. The demographic gradient between dense high-throughput populations where CIV is enzootic and smaller more diffuse populations where sporadic outbreaks may occur creates hotspot dynamics that can facilitate pathogen evolution toward higher transmissibility [25, 142], underscoring the potential role of urbanization in elevating the risk posed by emerging infectious pathogens [177, 135, 63].

Although humans exposed to these viruses appear not to be commonly infected (shown by serological testing) as the risk for human infection by either the EIV or CIV are currently unknown, as we do not understand the host barriers that restrict human infection, or the changes in the viruses that might allow those barriers to be overcome. Here we therefore provide a strategy for the preemptive eradication of an influenza A virus that is well adapted to mammals, since if CIV did gain high transmissibility among household dogs much of the human population would be directly exposed to the virus.

5.5 Acknowledgements

This chapter is based on the manuscript “Population dynamics, evolution, and control of emerging canine influenza virus in the United States” by Benjamin D. Dalziel, Kai Huang, Jemma L. Geoghegan, Nimalan Arinaminpathy, Edward J. Dubovi, Bryan T. Grenfell, Stephen P. Ellner, Edward C. Holmes and Colin R. Parrish. BDD, EJD, ECH and CRP conceived of the study. BDD collected the demographic data. BDD, NA, BTG and SPE designed the epidemic model. BDD and SPE analyzed the demographic data and epidemic model. KH did the laboratory and database work to assemble the genetic sequence data, and constructed the phylogenetic trees. JLG and ECH did the phylodynamic and phylogeographic analysis. BDD, JLG, SPE, ECH and CRP wrote the manuscript.

APPENDIX A

SUPPLEMENTARY INFORMATION FOR CHAPTER THREE

A.1 Kernel-smoothing model

\tilde{v}_i and \hat{v}_i are calculated as weighted averages of the actual caribou velocities in the data:

$$\tilde{v}_i = \sum_{j \in \tilde{\Omega}} p_{ij} v_j \quad (\text{A.1})$$

and

$$\hat{v}_i = \sum_{j \in \hat{\Omega}} p_{ij} v_j \quad (\text{A.2})$$

where p_{ij} represents the weights for a given kernel and bandwidths. We used a tricubic function to determine the weights:

$$p_{ij} \propto \begin{cases} (1 - |d_{ij}|^3)^3, & j \in \Omega_i \\ 0 & \text{otherwise} \end{cases} \quad (\text{A.3})$$

where the weights are normalized to satisfy $\sum_j p_{ij} = 1$. We used a tricubic function because it has compact support, is relatively flat for most of its support and is differentiable at the boundaries of its support [92]. A principal benefit of this kernel is that it thus avoids little contributions from extremely distant caribou. An alternate method for estimating advection fields from observations on the movements of particles is given by [178].

We found the best kernel bandwidths given the data using a search with cross-validation. The target was to maximize

$$R^2 = 1 - \frac{\sum_i |v_i - \hat{v}_i|^2}{\sum_i |v_i - \bar{v}|^2} \quad (\text{A.4})$$

as a function of κ and τ (which determine the predicted velocity \hat{v}_i), where \bar{v} is the global average velocity. This was ‘leave 10% out cross-validation’ where, for a given set of parameters, we randomized the order of the data, divided into 10 chunks, and used each possible set of 9 chunks to predict the one that was left out. This resulted in 10 R^2 values, which we then averaged to get the performance of the model for that set of parameters. Searching to maximize R^2 was done using the subplex algorithm [179, 180].

The focal observation i was not a member of any of its associated neighborhoods. Where the neighborhood would have been empty, it was populated with one randomly selected velocity from the entire dataset. Thus models with increasing numbers of empty neighborhoods converge to $R^2 = 0$ under cross-validation.

Particle simulation

Time step for the particle simulations was set to $h = 5$ days, matching the caribou data. At each time step in the simulations, the velocity of the particles was calculated from the best-fitting advection diffusion model. The positions of the particles were then updated using Equation (3.2).

A.2 Properties of ψ for ensembles of random walkers

We begin with a group of N independent particles in a one-dimensional environment, whose position at time t is given by

$$X_{i,t} = \varepsilon t + \delta W_{i,t} \quad (\text{A.5})$$

where $X_{i,t} \in \mathbb{R}$ is the location of the i th particle at time t and $W_{i,t}$ represents a Weiner process. The drift parameter ε represents the strength of the advection field and δ controls the diffusion rate of the particles.

Particles are observed once at time 0 and then again at time t . Define the velocity of the i th particle as

$$V_{i,t} = X_{i,t} - X_{i,0} \quad (\text{A.6})$$

and the normalized velocity (direction) as

$$U_{i,t} = \frac{V_{i,t}}{|V_{i,t}|} \quad (\text{A.7})$$

where $|\cdot|$ represents the vector norm.

The order statistic ψ is defined as

$$\psi_t = \frac{1}{N} \left| \sum_{i=1}^N U_{i,t} \right| \quad (\text{A.8})$$

Assume without loss of generality that $X_{i,0} = 0 \forall i$. Dropping the t subscript

$$\sum_{i=1}^N U_i = \sum_{i=1}^N R_i - \sum_{i=1}^N L_i \quad (\text{A.9})$$

where the indicator function R_i (or L_i) are equal to one if the i th individual is headed to the right (or left), and zero otherwise.

The probability that an individual is heading to the right (in the direction of the advection field), is determined by the strength of the advection field, ε , and the diffusion rate, which is proportional to δ .

$$p = \mathbf{Pr} [V_i > 0] \quad (\text{A.10})$$

$$= \mathbf{Pr} [X_i > 0] \quad (\text{A.11})$$

$$= 1 - \Phi(0) \quad (\text{A.12})$$

where $\Phi(\cdot)$ is the cumulative distribution function for a normal distribution with mean εt and variance $\delta^2 t$.

For sufficiently large N , $\sum_{i=1}^N U_i$ is normally distributed with mean $\mu = (2p - 1)N$ and variance $\sigma^2 = 4p(1 - p)N$.

ψ is then distributed according to a folded normal distribution. Define

$$\lambda = \sigma \sqrt{\frac{2}{\pi}} \exp\left(-\frac{\mu^2}{2\sigma^2}\right) + \mu \left(1 - 2\Phi\left(-\frac{\mu}{\sigma}\right)\right) \quad (\text{A.13})$$

where $\Phi(\cdot)$ here represents the cumulative distribution function of a standard normal distribution with mean zero and variance one, in contrast to its parameterization above,

Then

$$\mathbf{E}[\psi] = \frac{1}{N} \mathbf{E} \left[\left| \sum_{i=1}^N U_i \right| \right] \quad (\text{A.14})$$

$$= \frac{\lambda}{N} \quad (\text{A.15})$$

and

$$\mathbf{Var}[\psi] = \frac{1}{N^2} \mathbf{E} \left[\left| \sum_{i=1}^N U_i \right|^2 \right] \quad (\text{A.16})$$

$$= \frac{1}{N^2} (\mu^2 + \sigma^2 - \lambda^2) \quad (\text{A.17})$$

APPENDIX B

SUPPLEMENTARY INFORMATION FOR CHAPTER FOUR

B.1 Transportation models

The gravity model predicts commuting flow between CTs i and j as proportional to the product of powers of the total population of the source and destination divided by some function of the distance between them

$$\langle T_{ij} \rangle_{\text{gravity}} = \alpha \frac{N_i^\beta N_j^\gamma}{f(r_{ij})} \quad (\text{B.1})$$

where $N_i \geq n_i$ is the total population size of CT i and r_{ij} is equal to one plus the distance between CTs i and j in km . We used two alternate forms for the distance weighting function $f(r) = r^\delta$ and $f(r) = \exp(\delta r)$.

The configuration model is given by

$$\langle T_{ij} \rangle_{\text{configuration}} = n_i p_{ij} \quad (\text{B.2})$$

$$= n_i \frac{n_j}{\sum_j n_j} \quad (\text{B.3})$$

and the radiation model by

$$\langle T_{ij} \rangle_{\text{radiation}} = n_i p_{ij} \quad (\text{B.4})$$

$$= n_i \frac{N_i N_j}{(N_i + s_{ij})(N_i + N_j + s_{ij})} \quad (\text{B.5})$$

where s_{ij} is the total population within the circle of radius r_{ij} with its center at i . Both the configuration and radiation models are parameter free, and so did not require fitting to the commuting data.

We fit the gravity models to the commuting data for each city by minimizing

$$R_{\text{gravity}}^2 = \frac{\sum_i \sum_j (T'_{ij} - \langle T'_{ij} \rangle)^2}{\sum_i \sum_j (T'_{ij} - \bar{T}'_{ij})^2} \quad (\text{B.6})$$

as a function of the α, β, γ and δ , where $T'_{ij} = \log(T_{ij} + 1)$ and \bar{T}'_{ij} is the mean T'_{ij} for a given city.

As with both the radiation and configuration models, the proportion of the variance in commuting flows explained by both versions of the gravity model varied systematical among cities (Figure B.1). As with the configuration model, commuting flows in smaller cities more closely matched the predictions of the gravity model.

B.2 Transmission model

To translate commuting patterns into epidemic predictions, we first used the commuting data to estimate contact frequencies among each of the $W^2 - W$ pairs of workers in each city. We assumed the unobserved trajectory of a given individual a , $x_a(t) \in \mathbb{R}^2$, was generated by a mixed Ornstein-Uhlenbeck (OU) process. The mixing occurred by a stochastic process whereby individuals commuted between home (h) and work (w) states, with each state characterized by an average location $\bar{x}_a(s)$, $s \in \{h, w\}$ unique to each individual. We think of these coordinates as the location of an individual's bed and workstation. These individual-specific average locations were sampled from the commuting data by first choosing a home and work CT for each individual, denoted i_a and j_a , as

$$(i_a, j_a) \sim T_{ij}/W \tag{B.7}$$

and then assigning random coordinates for $\bar{x}_a(s)$ from within the respective CTs. Thus, as the number of individuals became large, the location of each individual's home and work were distributed identically to the commuting data.

Let $s_a(t)$ represent a 's state (i.e. either at home or at work) at time t . The actual trajectories are then seen as excursions from home or work locations

$$x_a(t) = \bar{x}_a(s_a(t)) + y(t) \quad (\text{B.8})$$

with the excursions, $y(t)$, independently and identically distributed for each individual and determined by a 2D OU process:

$$\frac{d}{dt}y(t) = -\mu y(t) + \sigma z(t) \quad (\text{B.9})$$

where the relaxation coefficient μ determines how rapidly individuals return from an excursion away from their state-specific average location (i.e. bed or workstation) and $\sigma z(t)$ represents a white noise process with total power σ . While individuals remain in a single state (i.e. either home or work), this yields a Gaussian stationary distribution for location with standard deviation

$$\rho = \left(\frac{\sigma^2}{2\mu}\right)^{1/2} \quad (\text{B.10})$$

Assuming the commuting process and excursions happen rapidly enough to become ergodic, it can be shown that the frequency with which two individuals a and b pass within an infective radius ε of each other is:

$$n_{ab} = \frac{\varepsilon^2}{4\rho^2} \sum_{s_a \in h,w} \sum_{s_b \in h,w} p_{s_a s_b} \exp\left(-\frac{\|\bar{x}_a(s_i) - \bar{x}_a(s_j)\|^2}{4\rho^2}\right) \quad (\text{B.11})$$

where $p_{s_a s_b}$ is the stationary distribution for the probability of finding two individuals in pairwise state $s_a s_b$, and $\varepsilon \ll \rho$.

We model the pathogen load of an infected individual t days after exposure using a basic model for acute infections[165]. In the model, pathogen load initially increases exponentially at rate α to peak at a time t^* , then declines ex-

ponentially at rate ω .

$$v(t) = v_0 \begin{cases} e^{\alpha(t-t^*)} & , t \leq t^* \\ e^{\omega(t-t^*)} & , t > t^* \end{cases} \quad (\text{B.12})$$

Following the appearance of one or more infected individuals, the rate of disease spread is determined by the matrix of random variables τ_{ab} , representing the time elapsed between when individual a becomes infected (if that happens) and when a first transmits the infection to b . We refer to $\tau = (\tau_{ab})$ as the transmission network. We assume that each individual can only be infected once, that a susceptible individual becomes infected the first time the pathogen is transmitted to them, and that subsequent transmissions do not affect their state. We let the distribution of τ_{ab} be determined by a time-varying transmission hazard proportional to $v(t)$ and to n_{ab} . This is a model for airborne pathogens transmitted by proximity between hosts.

Standard hazard analysis gives the distribution of τ_{ab} as

$$\text{P}[\tau_{ab} \leq t] = 1 - \exp\left(-\lambda n_{ab} \int_0^t v(t) dt\right) \quad (\text{B.13})$$

where λ converts units of pathogen load to transmission probability. For each city and transmissibility level we drew two replicate values for each τ_{ab} . Note that

$$\lambda n_{ab} = \lambda' \sum_{s_a \in h,w} \sum_{s_b \in h,w} p_{s_a s_b} \exp\left(-\frac{\|\bar{x}_a(s_i) - \bar{x}_a(s_j)\|^2}{4\rho^2}\right) \quad (\text{B.14})$$

where $\lambda' = \lambda \frac{\varepsilon^2}{4\rho^2}$. Thus in our model varying the transmissibility λ is equivalent to varying the ratio of the infectious radius to the radius of gyration of a host around their state-specific average location.

We used the following parameters: $\alpha = 2 \text{ days}^{-1}$, $\omega = -0.6 \text{ days}^{-1}$, $t^* = 2 \text{ days}$,

$\varepsilon = 10$, $m, \rho = 100$, $p_{s_a s_b} = (0.5, 0.1, 0.1, 0.3)$ for pairwise states (hh, hw, wh, ww).

We controlled for the effect of the topology of the transmission network given by τ on epidemic dynamics by running each simulation on three different types of transmission networks: the “real” one, generated from the commuting data; a transmission network constructed with the same data but with correlations between home and work locations broken; and a network with the same number of secondary transmissions for each individual as the real network, but where contact identity was randomized, assembled using a configuration model approach. The network made with the configuration model had a higher peak attack rate—peak attack rate is defined as the maximum proportion of the simulated population newly infected in a single day—than the network generated by the actual commuting patterns (Figure B.2). This shows that the main way mobility patterns affect epidemic dynamics in the simulations is by changing the degree distribution of the contact network. Peak attack rate is also affected by the fact that high-degree individuals are clustered in particular CTs and therefore preferentially link to each other; this feature is present in the “real” and “no home-work correlations” networks, but not in the configuration model.

When analyzing the simulation results we excluded simulation runs where the final epidemic size was < 5000 , since these corresponded to small outbreaks whose size was homogeneously random, and did not vary with N or m^* . For the configuration model network, we calculated as a reference the basic reproductive number of the simulated disease, representing the expected number of secondary infections produced by the first infected individual in the population, as $R_0 = \bar{k} + \sigma_k^2/\bar{k} - 1$ where \bar{k} and σ_k are the mean and standard deviation in the number of secondary infections produced by a randomly chosen infected

individual[181]. Closed formulae for R_0 are not currently available for more topologically complex networks (Figure B.3).

Table B.1: Summary of cities used in our study

City	N	no. CTs	Longitude	Latitude
Abbotsford	158799	34	-122.30	49.08
Barrie	177061	32	-79.68	44.37
Belleville	91432	28	-77.47	44.16
Brantford	124607	27	-80.28	43.16
Calgary	1079310	202	-114.08	51.05
Chilliwack	80872	28	-121.92	49.16
Drummondville	78108	17	-72.48	45.88
Edmonton	1027636	217	-113.52	53.54
Fredericton	85688	25	-66.67	45.97
Granby	68318	17	-72.73	45.39
Greater Sudbury	158258	41	-81.01	46.53
Guelph	127009	29	-80.25	43.54
Halifax	372858	87	-63.58	44.69
Hamilton	690369	176	-79.84	43.26
Kamloops	91214	26	-120.33	50.70
Kelowna	128588	28	-119.48	49.88
Kingston	122669	33	-76.53	44.27
Kitchener	450110	91	-80.45	43.43
Lethbridge	95091	25	-112.84	49.73
London	457720	103	-81.25	42.95
Medicine Hat	68822	16	-110.64	50.03
Moncton	126424	27	-64.81	46.09

Continued on next page

Table B.1 – *Continued from previous page*

City	N	no. CTs	Longitude	Latitude
Montreal	2895225	705	-73.63	45.54
Nanaimo	92361	20	-123.98	49.19
North Bay	63424	20	-79.42	46.30
Oshawa	245039	54	-78.85	43.91
Ottawa - Gatineau	902616	200	-75.69	45.40
Peterborough	107167	25	-78.32	44.31
Prince George	83003	26	-122.74	53.91
Quebec	538883	130	-71.26	46.82
Red Deer	82772	16	-113.80	52.27
Regina	194971	52	-104.62	50.46
Saguenay	151643	37	-71.12	48.41
Saint John	122374	45	-66.06	45.31
Saint-Jean-sur-Richelieu	87449	31	-73.28	45.32
Sarnia	87937	23	-82.37	42.97
Saskatoon	233893	51	-106.64	52.13
Sault Ste. Marie	80098	23	-84.31	46.53
Sherbrooke	161860	35	-71.92	45.38
St. Catharines - Niagara	390317	93	-79.20	43.08
St. John's	134902	36	-52.78	47.54
Thunder Bay	122907	33	-89.25	48.41
Toronto	3736916	743	-79.47	43.72
Trois-Rivieres	121068	32	-72.54	46.36
Vancouver	1615843	312	-122.97	49.22

Continued on next page

Table B.1 – *Continued from previous page*

City	N	no. CTs	Longitude	Latitude
Victoria	284495	60	-123.39	48.46
Windsor	323342	70	-82.98	42.28
Winnipeg	694668	167	-97.14	49.89

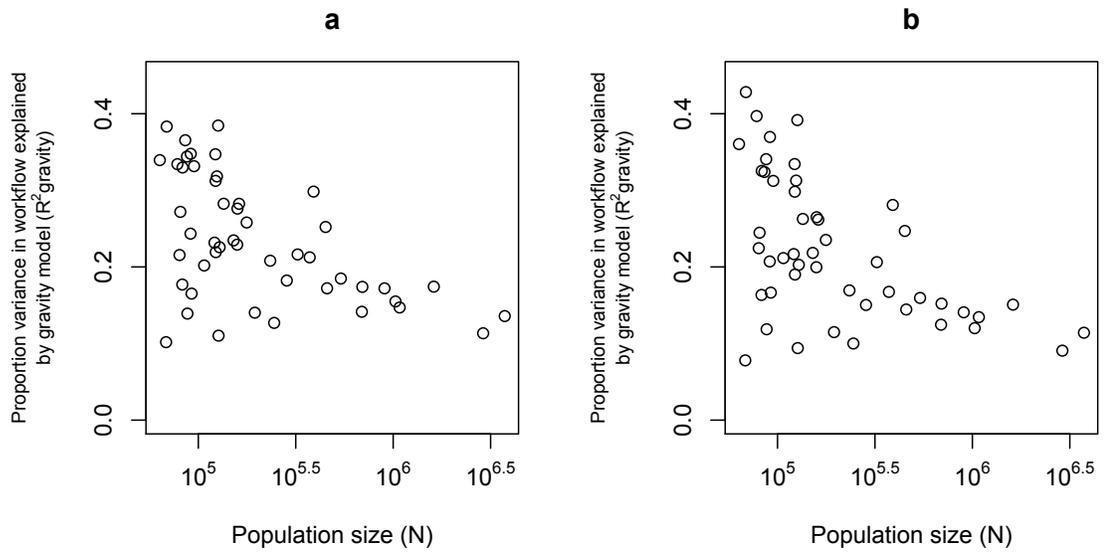


Figure B.1: **Fit of the gravity model in each city as a function of population size.** We used alternate distance weighting functions ($f(r)$) in the denominator. a: shows $f(r) = r^{-\delta}$. b: shows $f(r) = \exp(-\delta r)$

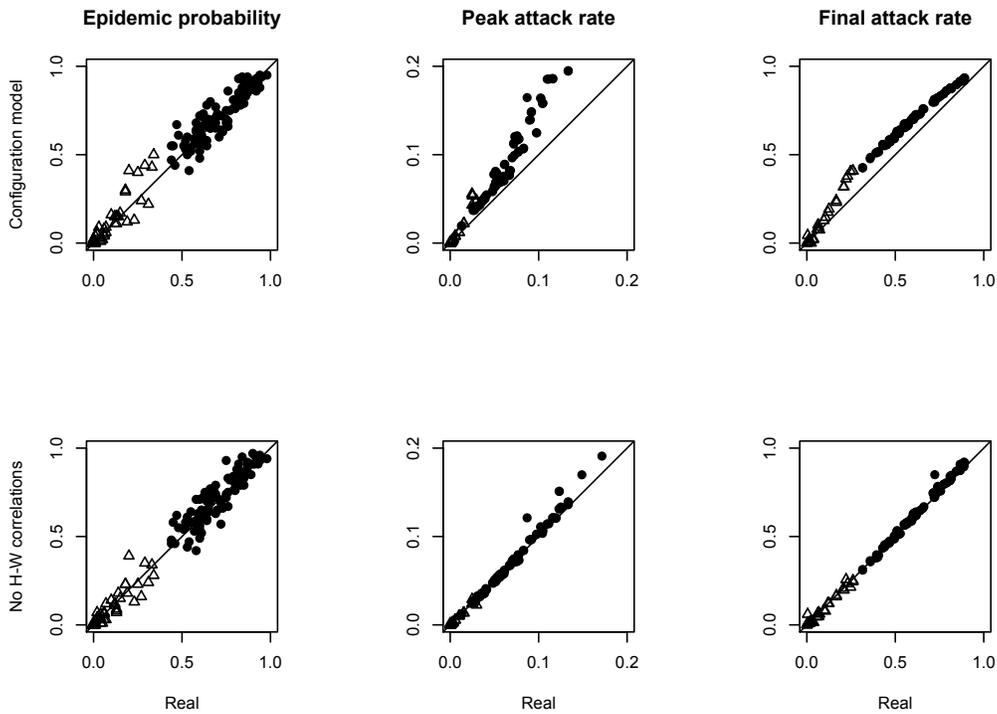


Figure B.2: Comparing the values of epidemic probes from the simulations—the probability that a single initial infection will spark an epidemic, peak attack rate and final attack rate—across different network topologies—real, no home-work correlations, and configuration model—for $\lambda = 1$ (triangles) and $\lambda = 10$ (circles).

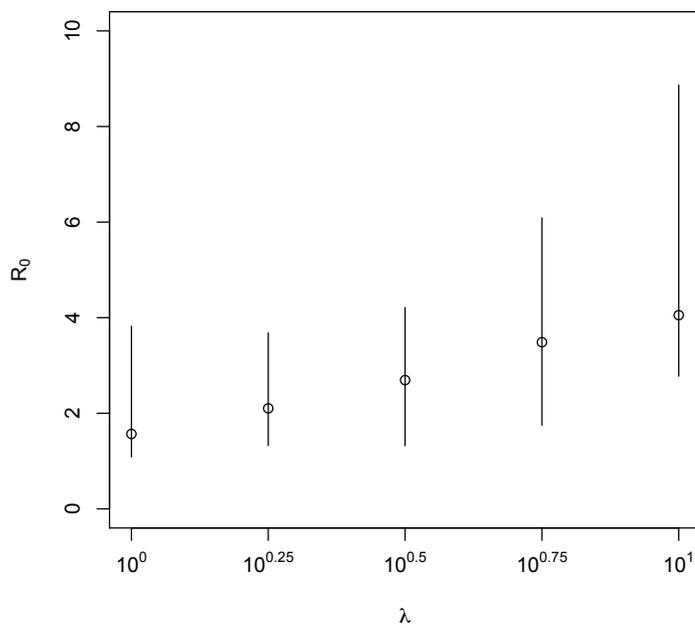


Figure B.3: R_0 on the configuration model network as a function of λ . Points show median values across cities and simulation runs; vertical lines go from the first quartile to the third quartile of the distribution across cities and simulation runs.

APPENDIX C
SUPPLEMENTARY INFORMATION FOR CHAPTER FIVE

C.1 Mean field model of a single shelter

$$\dot{S} = \mu - (\lambda + \alpha + \delta)S \quad (\text{C.1})$$

$$\dot{I} = \lambda S - (\gamma + \delta)I \quad (\text{C.2})$$

$$\dot{R} = \gamma(I + W) - \delta R \quad (\text{C.3})$$

$$\dot{V} = \alpha S - (\varepsilon\lambda + \delta)V \quad (\text{C.4})$$

$$\dot{W} = \varepsilon\lambda V - (\gamma + \delta)W \quad (\text{C.5})$$

where

$$\lambda = \frac{\beta(I + \omega W)}{N} \quad (\text{C.6})$$

The base model without vaccination is a special case where $\alpha = 0$, and $V = W = 0$

Disease-free equilibrium

$$S = \frac{\mu}{\alpha + \delta} \quad (\text{C.7})$$

$$V = \frac{\alpha}{\alpha + \delta} N \quad (\text{C.8})$$

$$(\text{C.9})$$

where $N = \mu/\delta$.

C.2 R_0 from mean field absent vaccination

We use the spectral radius method [182]. Let x represent the state vector for the system, so that the i th element of x corresponds to the value of the i th state variable. Let x_0 represent the disease free state. Let $\mathcal{F}_i(x)$ be the rate of appearance of new infected individuals into class i and $\mathcal{V}_i(x)$ represent the rate at which infected individuals leave that class. Define the matrices $\mathbf{F} = [\mathbf{F}_{ij}]$ and $\mathbf{V} = [\mathbf{V}_{ij}]$ such that

$$\mathbf{F}_{ij} = \frac{\partial \mathcal{F}_i}{\partial x_j}(x_0) \quad (\text{C.10})$$

and

$$\mathbf{V}_{ij} = \frac{\partial \mathcal{V}_i}{\partial x_j}(x_0) \quad (\text{C.11})$$

where i and j cover only the classes of infected individuals. Then

$$R_0 = \rho(\mathbf{FV}^{-1}) \quad (\text{C.12})$$

where ρ is the spectral radius of the resulting matrix.

Absent vaccination, we have

$$\mathcal{F} = \frac{\beta}{N}SI \quad (\text{C.13})$$

and

$$\mathcal{V} = (\gamma + \delta)I \quad (\text{C.14})$$

This leads to

$$R_0 = \frac{\beta}{\gamma + \delta} \quad (\text{C.15})$$

R_0 from mean field, full model

Vaccine coverage, C , at the disease free equilibrium is

$$C \equiv \frac{V}{N} = 1 - \frac{S}{N} = \frac{\alpha}{\alpha + \delta} \quad (\text{C.16})$$

Then

$$\mathbf{F} = \beta \begin{bmatrix} 1 - C & \omega(1 - C) \\ \varepsilon C & \varepsilon \omega C \end{bmatrix} \quad (\text{C.17})$$

and

$$\mathbf{V} = (\gamma + \delta) \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \quad (\text{C.18})$$

so

$$\mathbf{FV}^{-1} = \frac{\beta}{\gamma + \delta} \begin{bmatrix} 1 - C & \omega(1 - C) \\ \varepsilon C & \varepsilon \omega C \end{bmatrix} \quad (\text{C.19})$$

Letting R_0^+ represent the R_0 under vaccination, we have

$$R_0^+ = \rho(\mathbf{FV}^{-1}) \quad (\text{C.20})$$

$$= \frac{\beta}{\gamma + \delta} (1 + \varepsilon \omega C - C) \quad (\text{C.21})$$

This can be made more familiar by substituting the vaccine-free R_0 in place of $\beta/(\gamma + \delta)$ and defining K as the effective coverage, adjusted for vaccine performance,

$$K \equiv (1 - \varepsilon \omega)C \quad (\text{C.22})$$

to yield

$$R_0^+ = R_0 (1 - K) \quad (\text{C.23})$$

Endemic equilibrium absent vaccination

Starting with the I-nullcline

$$\dot{I} = 0 \quad (\text{C.24})$$

$$\iff S = \frac{\gamma + \delta}{\beta} N \quad (\text{C.25})$$

$$\iff \frac{S}{N} = \frac{1}{R_0} \quad (\text{C.26})$$

Then the S-nullcline

$$\dot{S} = 0 \quad (\text{C.27})$$

$$\iff I = \frac{\mu N - \delta S N}{\beta S} \quad (\text{C.28})$$

$$= \frac{\mu N}{\beta S} - \frac{\delta}{\beta} N \quad (\text{C.29})$$

Substituting in $N/S = R_0$, the S-nullcline satisfies

$$I = \frac{\mu}{\beta} R_0 - \frac{\delta}{\beta} N \quad (\text{C.30})$$

Prevalence

Prevalence at endemic equilibrium is given by

$$P \equiv \frac{I}{N} \quad (\text{C.31})$$

$$= \frac{\delta}{\beta} (R_0 - 1) \quad (\text{C.32})$$

$$= \frac{\delta}{\gamma + \delta} - \frac{\delta}{\beta} \quad (\text{C.33})$$

$$= \frac{\delta}{\gamma + \delta} \left(1 - \frac{1}{R_0} \right) \quad (\text{C.34})$$

C.3 Seroprevalence

Time to seroconversion is close to duration of viral shedding at around 7 days, so we assume recovered individuals have seroconverted and infected individuals have not [149]. Equilibrium (long-term average) seroprevalence in a shelter where CIV is endemic is thus given by

$$Z^* \equiv \frac{R}{N} \quad (\text{C.35})$$

$$= 1 - P - S/N \quad (\text{C.36})$$

$$= 1 - \frac{\delta}{\gamma + \delta} \left(1 - \frac{1}{R_0}\right) - \frac{1}{R_0} \quad (\text{C.37})$$

C.4 Metapopulation model

We have individual level data for 124,519 dogs from 13 animal shelters, spanning 2008 to 2013. Each row in the resulting data frame consists of the arrival and departure dates of one dog from a specified shelter. The data also contained a column giving information on outcome type, allowing us to exclude dogs that were admitted as the result of euthanasia requests, as their residence times were systematically lower than that over other dogs. In the i th shelter we estimate the arrival rate μ_i as the median number of dogs arriving per day to that shelter over the duration of the observation period. The departure rate in shelter i is estimated by taking the inverse of the median length of stay for dogs in that shelter over the duration of the observation period. Equilibrium shelter population size for shelter i is then given as $N_i = \mu_i/\delta_i$. We reconstructed the shelter's actual population size over time by subtracting cumulative departures from cumulative arrivals, and visually checked the stationarity of each shelter's

actual population size against the equilibrium value given by N_i . Eight shelters recorded whether outcomes were transfers to other shelters. In each of these eight shelters we calculated the proportion of all outcomes that were transfers. The average of this value across shelters is $\tau = 0.1$ (0.1 s.d). When estimating intake, output, and transfer we excluded dogs whose length of stay was greater than 40 days, as these represented rare atypical cases.

To parameterize the metapopulation model we need the arrival rate of dogs to each shelter from each other shelter μ_{ij} . We also need the per-capita rate at which shelter i transfers dogs to shelter j , δ_{ij} . Let μ_{ii} represent arrivals that are not transfers from another shelter (e.g. owner surrender) and let δ_{ii} represent the per-capita rate of outcomes that are not transfers to another shelter (e.g., adoption, euthanasia).

Here is how we go from the demographic data we have to the metapopulation parameters we need. The mean per capita departure rate for the shelter associated with a randomly chosen dog (which we have) satisfies

$$\delta_i = \sum_{j=1}^M \delta_{ij} \quad (\text{C.38})$$

where there are M shelters in the metapopulation.

A proportion τ of these outcomes are transfers to other shelters, and we have an estimate of τ . Assume that shelters accept dogs transferred from other shelters in the metapopulation proportional to the size of the recipient shelter, so that large shelters accept more transfers than small ones. For shelter i , define the proportion of the metapopulation outside of shelter i that resides in shelter j as

$$p_{ij} = \frac{N_j}{\sum_{k \neq i} N_k}, j \neq i \quad (\text{C.39})$$

which we can also calculate from the demographic data we have.

Assuming the metapopulation is closed, so that no dogs are transferred to shelters not in the metapopulation, we get δ_{ij} as

$$\delta_{ij} = \tau \delta_i p_{ij}, j \neq i \quad (\text{C.40})$$

Substituting into equation C.38 yields,

$$\delta_{ii} = \left(1 - \tau \sum_{j=1}^M p_{ij} \right) \delta_i \quad (\text{C.41})$$

Now let μ_{ij} represent the number of dogs that arrive to shelter i from j per day. Then

$$\mu_{ij} = \delta_{ji} N_j \quad (\text{C.42})$$

We assume that during transfers dogs are selected at random from the shelter without regard to their disease state. Thus the probability that a susceptible dog is transferred is S_i/N_i and so on for the other classes.

Finally, we choose μ_{ii} to balance the relation

$$\mu_i = \sum_j \mu_{ij} \quad (\text{C.43})$$

where the left-hand side is given by data.

In contrast to μ_{ij} , the flow represented by μ_{ii} is assumed to consist of entirely susceptible dogs, because the prevalence of CIV among dogs not in animal shelters is very low.

Our metapopulation approach ignores shelter proximity when calculating transfer probabilities, because we have no data on how transfer rates vary with

geographic distance. We therefore use the mean transfer rate τ as an appropriately simple first approximation for transfer dynamics among shelters. The relative low mean transfer probability of $\tau = 0.1$ suggests that transfer of dogs among shelters may be infrequent enough to prevent panmixia. That is, on a continuum from completely isolated to completely intermixed, dog populations in animal shelters in the US are closer to being solitary than to being totally connected. This is broadly consistent with finding of geographic segregation in the phylogenetic data.

BIBLIOGRAPHY

- [1] Malthus TR (1798) An essay on the principle of population. Printed for J Johnson, In St Paul's Church Yard .
- [2] Verhulst PF (1845) Recherches mathématiques sur la loi d'accroissement de la population. Nouv mém de l'Academie Royale des Sci et Belles-Lettres de Bruxelles 18: 1–41.
- [3] Lotka AJ (1920) Undamped oscillations derived from the law of mass action. J Am Chem Soc 42: 1595–1599.
- [4] Volterra V (1926) Fluctuations in the abundance of a species considered mathematically. Nature 118: 558–560.
- [5] Kermack WO, McKendrick AG (1927) A contribution to the mathematical theory of epidemics. Proc R Soc A 115: 700–721.
- [6] Holling CS (1959) Some Characteristics of Simple Types of Predation and Parasitism. Can Entomol 91: 385–398.
- [7] Rosenzweig ML, MacArthur RH (1963) Graphical representation and stability conditions of predator-prey interactions. Am Nat 97: 209–223.
- [8] Waage P, Gulberg CM (1986) Studies concerning affinity. J Chem Educ 63: 1044.
- [9] Ovaskainen O, Cornell SJ (2006) Space and stochasticity in population dynamics. PNAS 103: 12781–12786.
- [10] Turchin P (2002) Does population ecology have general laws? Oikos 63: 3–14.
- [11] Ōkubo A, Levin S (2001) Diffusion and ecological problems: modern perspectives. Springer.
- [12] Fisher RA (1937) The wave of advance of advantageous genes. Ann Eugenetic 7: 355–369.
- [13] Skellam JG (1951) Random dispersal in theoretical populations. Biometrika 38: 196–218.

- [14] Caswell H (2001) Matrix population models. construction, analysis, and interpretation. Sinauer Associates Inc.
- [15] Ellner SP, Rees M (2006) Integral projection models for species with complex demography. *Am Nat* 167: 410–428.
- [16] Yoshida T, Jones L, Ellner S, Fussmann G, Hairston N (2003) Rapid evolution drives ecological dynamics in a predator–prey system. *Nature* 424: 303–306.
- [17] Koelle K, Cobey S, Grenfell B, Pascual M (2006) Epochal evolution shapes the phylodynamics of interpandemic influenza A (H3N2) in humans. *Science* 314: 1898–1903.
- [18] Schrodinger E (1944) What is life? Cambridge University Press.
- [19] Pascual M, Roy M, Laneri K (2011) Simple models for complex systems: exploiting the relationship between local and global densities. *Theor Ecol* 4: 211–222.
- [20] Machta BB, Chachra R, Transtrum MK, Sethna JP (2013) Parameter space compression underlies emergent theories and predictive models. *Science* 342: 604–607.
- [21] May R, Gupta S, McLean AR (2001) Infectious disease dynamics: what characterizes a successful invader? *Phil Trans R Soc Lond B* 356: 901–910.
- [22] Newman MEJ (2002) Spread of epidemic disease on networks. *Phys Rev E Stat Nonlin Soft Matter Phys* 66: 016128.
- [23] Lloyd-Smith J, Schreiber SJ, Kopp PE, Getz W (2005) Superspreading and the effect of individual variation on disease emergence. *Nature* 438: 355–359.
- [24] Bansal S, Grenfell BT, Meyers LA (2007) When individual behaviour matters: homogeneous and network models in epidemiology. *J R Soc Interface* 4: 879–891.
- [25] Lieberman E, Hauert C, Nowak M (2005) Evolutionary dynamics on graphs. *Nature* 433: 312–316.

- [26] Prigogine I (1978) Time, structure, and fluctuations. *Science* 201: 777–785.
- [27] Messier F (1994) Ungulate population models with predation: a case study with the North American Moose. *Ecology* 75: 478–488.
- [28] Grenfell B, Bjornstad O, Kappey J (2001) Travelling waves and spatial hierarchies in measles epidemics. *Nature* 414: 716–723.
- [29] Vicsek T (2001) A question of scale. *Nature* 411: 421.
- [30] Couzin ID, Krause J, Franks NR, Levin SA (2005) Effective leadership and decision-making in animal groups on the move. *Nature* 433: 513–516.
- [31] Saigusa T, Tero A, Nakagaki T, Kuramoto Y (2008) Amoebae anticipate periodic events. *Phys Rev Lett* 100: 018101.
- [32] Ginsberg J, Mohebbi MH, Patel RS, Brammer L, Smolinski MS, et al. (2009) Detecting influenza epidemics using search engine query data. *Nature* 457: 1012–1014.
- [33] Vicsek T, Czirók A, Ben-Jacob E, Cohen I, Shochet O (1995) Novel Type of Phase Transition in a System of Self-Driven Particles. *Phys Rev Lett* 75: 1226–1229.
- [34] Berdahl A, Torney CJ, Ioannou CC, Faria JJ, Couzin ID (2013) Emergent sensing of complex environments by mobile animal groups. *Science* 339: 574–576.
- [35] Sumpter DJT, Mann RP, Perna A (2012) The modelling cycle for collective animal behaviour. *Interface Focus* 2: 764–773.
- [36] Couzin ID, Krause J, James R, Ruxton GD, Franks NR (2002) Collective memory and spatial sorting in animal groups. *J Theor Biol* 218: 1–11.
- [37] Vicsek T, Zafeiris A (2012) Collective motion. *Phys Rep* 517: 71–140.
- [38] Grünbaum D (1998) Schooling as a strategy for taxis in a noisy environment. *Evol Ecol* 12: 503–522.
- [39] Simons AM (2004) Many wrongs: the advantage of group navigation. *Trends Ecol Evol* 19: 453–455.

- [40] Codling EA, Pitchford JW, Simpson SD (2007) Group navigation and the "many-wrongs principle" in models of animal movement. *Ecology* 88: 1864–1870.
- [41] Torney C, Neufeld Z, Couzin ID (2009) Context-dependent interaction leads to emergent search behavior in social aggregates. *PNAS* 106: 22055–22060.
- [42] Clark CW, Mangel M (1986) The evolutionary advantages of group foraging. *Theor Popul Biol* 30: 45–75.
- [43] Handegard NO, Boswell KM, Ioannou CC, Leblanc SP, Tjøstheim DB, et al. (2012) The Dynamics of Coordinated Group Hunting and Collective Information Transfer among Schooling Prey. *Curr Biol* 22: 1213–1217.
- [44] Ioannou CC, Guttal V, Couzin I (2012) Predatory Fish Select for Coordinated Collective Motion in Virtual Prey. *Science* 337: 1212–1215.
- [45] Olson RS, Hintze A, Dyer FC, Knoester DB, Adami C (2013) Predator confusion is sufficient to evolve swarming behaviour. *J R Soc Interface* 10: 20130305–20130305.
- [46] Nagy M, Akos Z, Biro D, Vicsek T (2010) Hierarchical group dynamics in pigeon flocks. *Nature* 464: 890–893.
- [47] Bode NWF, Franks DW, Wood AJ, Piercy JJB, Croft DP, et al. (2012) Distinguishing Social from Nonsocial Navigation in Moving Animal Groups. *Am Nat* 179: 621–632.
- [48] Gautrais J, Ginelli F, Fournier R, Blanco S, Soria M, et al. (2012) Deciphering interactions in moving animal groups. *PLoS Comput Biol* 8: e1002678.
- [49] Nagy M, Vásárhelyi G, Pettit B, Roberts-Mariani I, Vicsek T, et al. (2013) Context-dependent hierarchies in pigeons. *PNAS* 110: 13049–13054.
- [50] Eriksson A, Nilsson Jacobi M, Nystrom J, Tunstrom K (2010) Determining interaction rules in animal swarms. *Behav Ecol* 21: 1106–1111.
- [51] Levin S (1992) The problem of pattern and scale in ecology: the Robert H. MacArthur award lecture. *Ecology* 73: 1943–1967.

- [52] Ihle T (2011) Kinetic theory of flocking: Derivation of hydrodynamic equations. *Phys Rev E Stat Nonlin Soft Matter Phys* 83: 030901.
- [53] Couzin I, Ioannou CC, Demirel G, Gross T, Torney CJ, et al. (2011) Uninformed Individuals Promote Democratic Consensus in Animal Groups. *Science* 334: 1578–1580.
- [54] Franks NR, Pratt SC, Mallon EB, Britton NF, Sumpter DJT (2002) Information flow, opinion polling and collective intelligence in house-hunting social insects. *Phil Trans R Soc Lond B* 357: 1567–1583.
- [55] Perony N, Tessone CJ, König B, Schweitzer F (2012) How random is social behaviour? Disentangling social complexity through the study of a wild house mouse population. *PLoS Comput Biol* 8: e1002786.
- [56] Delgado MM, Delgado MdM, Penteriani V, Penteriani V, Morales JM, et al. (2014) A statistical framework for inferring the influence of conspecifics on movement behaviour. *Methods Ecol Evol* 5: 1–7.
- [57] Patlak CS (1953) Random walk with persistence and external bias. *Bull Math Biol* 15: 311–338.
- [58] Mitbavkar S, Anil AC (2004) Vertical migratory rhythms of benthic diatoms in a tropical intertidal sand flat: influence of irradiance and tides. *Marine Biology* 145.
- [59] Jorgensen HBH, Hansen MM, Bekkevold D, Ruzzante DE, Loeschcke V (2005) Marine landscapes and population genetic structure of herring (*Clupea harengus* L.) in the Baltic Sea. *Mol Ecol* 14: 3219–3234.
- [60] Huse G (2002) Modelling changes in migration pattern of herring: collective behaviour and numerical domination. *J Fish Biol* 60: 571–582.
- [61] Bettencourt LMA, Lobo J, Helbing D, Kühnert C, West GB (2007) Growth, innovation, scaling, and the pace of life in cities. *PNAS* 104: 7301–7306.
- [62] Batty M (2008) The size, scale, and shape of cities. *Science* 319: 769–771.
- [63] Dalziel BD, Pourbohloul B, Ellner S (2013) Human mobility patterns predict divergent epidemic dynamics among cities. *Proc R Soc B* 280: 20130763–20130763.

- [64] Daniels R, Vanderleyden J, Michiels J (2004) Quorum sensing and swarming migration in bacteria. *FEMS Microbiol Rev* 28: 261–289.
- [65] Chapman JW, Nesbit RL, Burgin LE, Reynolds DR, Smith AD, et al. (2006) Flight orientation behaviors promote optimal migration trajectories in high-flying insects. *Science* 313: 794–796.
- [66] Alerstam T (1990) Bird migration. Cambridge University Press.
- [67] Holdo RM, Holt RD, Fryxell JM (2009) Opposing rainfall and plant nutritional gradients best explain the wildebeest migration in the Serengeti. *Am Nat* 173: 431–445.
- [68] Alerstam T (2006) Conflicting Evidence About Long-Distance Animal Navigation. *Science* 313: 791–794.
- [69] Dunlap JC, Loros JJ, DeCoursey PJ (2004) Chronobiology. Biological Time-keeping. Sinauer Associates Incorporated.
- [70] Hein CM, Engels S, Kishkinev D, Mouritsen H (2011) Robins have a magnetic compass in both eyes. *Nature* 471: E11–2– discussion E12–3.
- [71] Alerstam T, Gudmundsson GA, Green M, Hedenström A (2001) Migration along orthodromic sun compass routes by arctic birds. *Science* 291: 300–303.
- [72] Alerstam T, Chapman JW, Bäckman J, Smith AD, Karlsson H, et al. (2011) Convergent patterns of long-distance nocturnal migration in noctuid moths and passerine birds. *Proc R Soc B* 278: 3074–3080.
- [73] Guttal V, Couzin ID (2010) Social interactions, information use, and the evolution of collective migration. *PNAS* 107: 16172–16177.
- [74] Torney CJ, Levin SA, Couzin ID (2010) Specialization and evolutionary branching within migratory populations. *PNAS* 107: 20394–20399.
- [75] Conradt L, Roper TJ (2003) Group decision-making in animals. *Nature* 421: 155–158.
- [76] Helm B, Piersma T, van der Jeugd H (2006) Sociable schedules: interplay between avian seasonal and social behaviour. *Anim Behav* 72: 245–262.

- [77] Gueron S, Levin SA (1993) Self-organization of Front Patterns in Large Wildebeest Herds. *J Theor Biol* 165: 541–552.
- [78] Haydon DT, Morales JM, Yott A, Jenkins DA, Rosatte R, et al. (2008) Socially informed random walks: incorporating group dynamics into models of population spread and growth. *Proc R Soc B* 275: 1101–1109.
- [79] Buhl J, Sumpter DJT, Couzin I, Hale JJ, Despland E, et al. (2006) From disorder to order in marching locusts. *Science* 312: 1402–1406.
- [80] Bazazi S, Buhl J, Hale JJ, Anstey ML, Sword GA, et al. (2008) Collective motion and cannibalism in locust migratory bands. *Curr Biol* 18: 735–739.
- [81] Ward AJW, Sumpter DJT, Couzin ID, Hart PJB, Krause J (2008) Quorum decision-making facilitates information transfer in fish shoals. *PNAS* 105: 6948–6953.
- [82] Schellinck J, White T (2011) A review of attraction and repulsion models of aggregation: Methods, findings and a discussion of model validation. *Ecol Model* 222: 1897–1911.
- [83] Le Henaff D (1976) Inventaire aérien des terrains de vèlage du caribou dans la région nord et au nord du territoire de la municipalité de la Baie James (mai-juin 1975). Quebec: Ministère du tourisme, de la chasse et de la pêche.
- [84] Couturier S, Otto RD, Côté SD, Luther G, Mahoney SP (2010) Body size variations in caribou ecotypes and relationships with demography. *J Wildl Manag* 74: 395–404.
- [85] Taillon J, Festa-Bianchet M, Côté SD (2012) Shifting targets in the tundra: Protection of migratory caribou calving grounds must account for spatial changes over time. *Biol Conserv* 147: 163–173.
- [86] Post E, Forchhammer MC (2008) Climate change reduces reproductive success of an Arctic herbivore through trophic mismatch. *Phil Trans R Soc Lond B* 363: 2369–2375.
- [87] Boulet M, Couturier S, Côté SD, Otto RD, Bernatchez L (2007) Integrative use of spatial, genetic, and demographic analyses for investigating genetic

- connectivity between migratory, montane, and sedentary caribou herds. *Mol Ecol* 16: 4223–4240.
- [88] Escudero C, Yates CA, Buhl J, Couzin ID, Erban R, et al. (2010) Ergodic directional switching in mobile insect groups. *Phys Rev E Stat Nonlin Soft Matter Phys* 82: 011926.
- [89] Bergman CM, Schaefer JA, Luttich SN (2000) Caribou movement as a correlated random walk. *Oecologia* 123: 364–374.
- [90] Viswanathan GM (2010) Ecology: Fish in Lévy-flight foraging. *Nature* 465: 1018–1019.
- [91] Morales JM, Haydon DT, Frair JL, Holsinger KE, Fryxell JM (2004) Extracting more out of relocation data: building movement models as mixtures of random walks. *Ecology* 85: 2436–2445.
- [92] Hastie T, Tibshirani R, Friedman J (2008) *The elements of statistical learning: data mining, inference, and prediction*. Springer, 2nd edition edition.
- [93] Vicsek T (1999) Application of statistical mechanics to collective motion in biology. *Physica A* 274: 182–189.
- [94] Czirók A, Stanley HE, Vicsek T (1999) Spontaneously ordered motion of self-propelled particles. *J Phys A: Math Gen* 30: 1375–1385.
- [95] Smith KG, Ficht EJ, Hobson D (2000) Winter distribution of woodland caribou in relation to clear-cut logging in west-central Alberta. *Can J Zool* 78: 1433–1440.
- [96] Wilcove DS, Wikelski M (2008) Going, going, gone: is animal migration disappearing. *PLoS Biol* 6: e188.
- [97] Mishra S, Tunstrøm K, Couzin ID, Huepe C (2012) Collective dynamics of self-propelled particles with variable speed. *Phys Rev E Stat Nonlin Soft Matter Phys* 86: 011901.
- [98] Bazazi S, Romanczuk P, Thomas S, Schimansky-Geier L, Hale JJ, et al. (2011) Nutritional state and collective motion: from individuals to mass migration. *Proc R Soc B* 278: 356–363.
- [99] World Health Organization (2008). *Causes of death 2008 summary tables*.

- [100] World Health Organization (2008). The global burden of disease: 2004 update.
- [101] Bartlett M (1956) Measles periodicity and community size. *J R Stat Soc A* 120: 48–70.
- [102] Meyers LA, Pourbohloul B, Newman M, Skowronski D, Brunham R (2005) Network theory and SARS: predicting outbreak diversity. *J Theor Biol* 232: 71–81.
- [103] Taylor C, Marathe A, Beckman R (2010) Same influenza vaccination strategies but different outcomes across US cities? *International Journal of Infectious Diseases* 14: e792–e795.
- [104] Merler S, Ajelli M (2010) The role of population heterogeneity and human mobility in the spread of pandemic influenza. *Proc R Soc B* 277: 557–565.
- [105] Salathé M, Kazandjieva M, Lee JW, Levis P, Feldman MW, et al. (2010) A high-resolution human contact network for infectious disease transmission. *PNAS* 107: 22020–22025.
- [106] Colizza V, Barrat A, Barthelemy M, Valleron AJ, Vespignani A (2007) Modeling the Worldwide Spread of Pandemic Influenza: Baseline Case and Containment Interventions. *PLoS Med* 4: e13.
- [107] Balcan D, Hu H, Goncalves B, Bajardi P, Poletto C, et al. (2009) Seasonal transmission potential and activity peaks of the new influenza A(H1N1): a Monte Carlo likelihood analysis based on human mobility. *BMC Med* 7: 45.
- [108] Bajardi P, Poletto C, Ramasco JJ, Tizzoni M, Colizza V, et al. (2011) Human mobility networks, travel restrictions, and the global spread of 2009 H1N1 pandemic. *PLoS ONE* 6: e16591.
- [109] González MC, Hidalgo CA, Barabási AL (2008) Understanding individual human mobility patterns. *Nature* 453: 779–782.
- [110] Song C, Qu Z, Blumm N, Barabási AL (2010) Limits of predictability in human mobility. *Science* 327: 1018–1021.
- [111] Viboud C, Bjørnstad ON, Smith DL, Simonsen L, Miller MA, et al. (2006)

Synchrony, waves, and spatial hierarchies in the spread of influenza. *Science* 312: 447–451.

- [112] Balcan D, Vespignani A (2011) Phase transitions in contagion processes mediated by recurrent mobility patterns. *Nat Phys* 7: 581–586.
- [113] Barabási AL (2005) The origin of bursts and heavy tails in human dynamics. *Nature* 435: 207–211.
- [114] Mossong J, Hens N, Jit M, Beutels P, Auranen K, et al. (2008) Social contacts and mixing patterns relevant to the spread of infectious diseases. *PLoS Med* 5: e74.
- [115] Eubank S, Guclu H, Kumar VSA, Marathe MV, Srinivasan A, et al. (2004) Modelling disease outbreaks in realistic urban social networks. *Nature* 429: 180–184.
- [116] Guzzetta G, Ajelli M, Yang Z, Merler S (2011) Modeling socio-demography to capture tuberculosis transmission dynamics in a low burden setting. *J Theor Biol* 289: 197–205.
- [117] Fumanelli L, Ajelli M, Manfredi P, Vespignani A, Merler S (2012) Inferring the Structure of Social Contacts from Demographic Data in the Analysis of Infectious Diseases Spread. *PLoS Comput Biol* 8: e1002673.
- [118] Colizza V, Vespignani A (2007) Invasion threshold in heterogeneous metapopulation networks. *Phys Rev Lett* 99: 148701.
- [119] Colizza V, Pastor-Satorras R, Vespignani A (2007) Reaction–diffusion processes and metapopulation models in heterogeneous networks. *Nat Phys* 3: 276–282.
- [120] Colizza V, Vespignani A (2008) Epidemic modeling in metapopulation systems with heterogeneous coupling pattern: Theory and simulations. *J Theor Biol* 251: 450–467.
- [121] Statistics Canada (2008). Census of Canada. 2006 Census Place of Work (POW) Custom Consortium Tables. <http://hdl.handle.net/10573/41533>.
- [122] Statistics Canada (2010) Census dictionary. Statistics Canada.
- [123] Lloyd M (1967) Mean crowding. *J Anim Ecol* 36: 1–30.

- [124] Simini F, González MC, Maritan A, Barabási AL (2012) A universal model for mobility and migration patterns. *Nature* 484: 96–100.
- [125] Jesse M, Ezanno P, Davis S, Heesterbeek J (2008) A fully coupled, mechanistic model for infectious disease dynamics in a metapopulation: Movement and epidemic duration. *J Theor Biol* 254: 331–338.
- [126] Chowell G, Bettencourt LMA, Johnson N, Alonso WJ, Viboud C (2008) The 1918-1919 influenza pandemic in England and Wales: spatial patterns in transmissibility and mortality impact. *Proc R Soc B* 275: 501–509.
- [127] Bharti N, Tatem AJ, Ferrari MJ, Grais RF, Djibo A, et al. (2011) Explaining seasonal fluctuations of measles in Niger using nighttime lights imagery. *Science* 334: 1424–1427.
- [128] Wesolowski A, Eagle N, Tatem AJ, Smith D, Noor AM, et al. (2012) Quantifying the Impact of Human Mobility on Malaria. *Science* 338: 267–270.
- [129] Keeling MJ, Danon L, Vernon MC, House TA (2010) Individual identity and movement networks for disease metapopulations. *PNAS* 107: 8866–8870.
- [130] Poletto C, Tizzoni M, Colizza V (2012) Heterogeneous length of stay of hosts' movements and spatial epidemic spread. *Sci Rep* 2: 476.
- [131] Grenfell BT, Bolker BM (1998) Cities and villages: infection hierarchies in a measles metapopulation. *Ecol Lett* 1: 63–70.
- [132] Eames KTD, Tilston NL, Brooks-Pollock E, Edmunds WJ (2012) Measured dynamic social contact patterns explain the spread of H1N1v influenza. *PLoS Comput Biol* 8: e1002425.
- [133] Tizzoni M, Bajardi P, Poletto C, Ramasco JJ, Balcan D, et al. (2012) Real-time numerical forecast of global epidemic spreading: case study of 2009 A/H1N1pdm. *BMC Med* 10: 165.
- [134] Parrish CR, Kawaoka Y (2005) The origins of new pandemic viruses: the acquisition of new host ranges by canine parvovirus and influenza A viruses. *Annu Rev Microbiol* 59: 553–586.
- [135] Woolhouse MEJ, Haydon DT, Antia R (2005) Emerging pathogens: the

epidemiology and evolution of species jumps. *Trends Ecol Evol* 20: 238–244.

- [136] Wolfe ND, Dunavan CP, Diamond J (2007) Origins of major human infectious diseases. *Nature* 447: 279–283.
- [137] Guan Y, Zheng BJ, He YQ, Liu XL, Zhuang ZX, et al. (2003) Isolation and characterization of viruses related to the SARS coronavirus from animals in southern China. *Science* 302: 276–278.
- [138] Smith G, Vijaykrishna D, Bahl J, Lycett S, Worobey M, et al. (2009) Origins and evolutionary genomics of the 2009 swine-origin H1N1 influenza A epidemic. *Nature* 459: 1122–1125.
- [139] Morse SS, Mazet JAK, Woolhouse MEJ, Parrish CR, Carrol D, et al. (2012) Prediction and prevention of the next pandemic zoonosis. *The Lancet* 380: 1956–1965.
- [140] Antia R, Regoes RR, Koella JC, Bergstrom CT (2003) The role of evolution in the emergence of infectious diseases. *Nature* 426: 658–661.
- [141] Lloyd-Smith JO, George D, Pepin KM, Pitzer VE, Pulliam JRC, et al. (2009) Epidemic dynamics at the human-animal interface. *Science* 326: 1362–1367.
- [142] Arinaminpathy N, McLean AR (2009) Evolution and emergence of novel human infections. *Proc R Soc B* 276: 3937–3943.
- [143] Stadler T, Kühnert D, Bonhoeffer S, Drummond AJ (2013) Birth–death skyline plot reveals temporal changes of epidemic spread in HIV and hepatitis C virus (HCV). *PNAS* 110: 228–233.
- [144] Crawford PC, Dubovi EJ, Castleman WL, Stephenson I, Gibbs EPJ, et al. (2005) Transmission of equine influenza virus to dogs. *Science* 310: 482–485.
- [145] Hayward JJ, Dubovi EJ, Scarlett JM, Janeczko S, Holmes EC, et al. (2010) Microevolution of canine influenza virus in shelters and its molecular epidemiology in the United States. *J Virol* 84: 12636–12645.
- [146] Anderson TC, Bromfield CR, Crawford PC, Dodds WJ, Gibbs EPJ, et al.

- (2012) Serological evidence of H3N8 canine influenza-like virus circulation in USA dogs prior to 2004. *Vet J* 191: 312–316.
- [147] Rivaille P, Perry IA, Jang Y, Davis CT, Chen LM, et al. (2010) Evolution of canine and equine influenza (H3N8) viruses co-circulating between 2005 and 2008. *Virology* 408: 71–79.
- [148] Barrell EA, Pecoraro HL, Torres-Henderson C, Morley PS, Lunn KF, et al. (2010) Seroprevalence and risk factors for canine H3N8 influenza virus exposure in household dogs in Colorado. *J Vet Intern Med* 24: 1524–1527.
- [149] Jirjis FF, Deshpande MS, Tubbs AL, Jayappa H, Lakshmanan N, et al. (2010) Transmission of canine influenza virus (H3N8) among susceptible dogs. *Vet Microbiol* 144: 303–309.
- [150] Serra VF, Stanzani G, Smith G, Otto CM (2011) Point seroprevalence of canine influenza virus H3N8 in dogs participating in a flyball tournament in Pennsylvania. *J Am Vet Med Assoc* 238: 726–730.
- [151] Dubovi EJ, Njaa BL (2008) Canine Influenza. *Veterinary Clinics of North America: Small Animal Practice* 38: 827–835.
- [152] Holt DE, Mover MR, Brown DC (2010) Serologic prevalence of antibodies against canine influenza virus (H3N8) in dogs in a metropolitan animal shelter. *J Am Vet Med Assoc* 237: 71–73.
- [153] Waddell GH, Teigland MB, Sigel MM (1963) A new influenza virus associated with equine respiratory disease. *J Am Vet Med Assoc* 143: 587–590.
- [154] Daly JM, MacRae S, Newton JR, Watrang E, Elton DM (2011) Equine influenza: a review of an unpredictable virus. *Vet J* 189: 7–14.
- [155] Virmani N, Bera BC, Singh BK, Shanmugasundaram K, Gulati BR, et al. (2010) Equine influenza outbreak in India (2008-09): virus isolation, sero-epidemiology and phylogenetic analysis of HA gene. *Vet Microbiol* 143: 224–237.
- [156] Wei G, Xue-Feng L, Yan Y, Ying-Yuan W, Ling-Li D, et al. (2010) Equine influenza viruses isolated during outbreaks in China in 2007 and 2008. *Vet Rec* 167: 382–383.

- [157] Bountouri M, Fragkiadaki E, Ntafis V, Kanellos T, Xylouri E (2011) Phylogenetic and molecular characterization of equine H3N8 influenza viruses from Greece (2003 and 2007): evidence for reassortment between evolutionary lineages. *Virology* 8: 350.
- [158] Cowled B, Ward MP, Hamilton S, Garner G (2009) The equine influenza epidemic in Australia: Spatial and temporal descriptive analyses of a large propagating epidemic. *Prev Vet Med* 92: 60–70.
- [159] Hughes J, Allen RC, Baguelin M, Hampson K, Baillie GJ, et al. (2012) Transmission of Equine Influenza Virus during an Outbreak Is Characterized by Frequent Mixed Infections and Loose Transmission Bottlenecks. *PLoS Pathog* 8: e1003081.
- [160] Daly JM, Lai ACK, Binns MM, Chambers TM, Barrandeguy M, et al. (1996) Antigenic and genetic evolution of equine H3N8 influenza A viruses. *Journal of General Virology* 77: 661–671.
- [161] Murcia PR, Baillie GJ, Stack JC, Jervis C, Elton D, et al. (2013) Evolution of equine influenza virus in vaccinated horses. *J Virol* 87: 4768–4771.
- [162] McCallum H, Barlow N, Hone J (2001) How should pathogen transmission be modelled? *Trends Ecol Evol* 16: 295–300.
- [163] Gillespie DT (1977) Exact stochastic simulation of coupled chemical reactions. *J Phys Chem* 81: 2340–2361.
- [164] Gilks WR, Richardson S, Spiegelhalter D (1995) *Markov Chain Monte Carlo in Practice*. CRC Press.
- [165] Nowak M, May R (2000) *Virus dynamics*. Oxford University Press.
- [166] Edgar RC (2004) MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* 5: 113.
- [167] Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, et al. (2010) New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst Biol* 59: 307–321.
- [168] Stadler T, Kouyos R, von Wyl V, Yerly S, Böni J, et al. (2012) Estimating the basic reproductive number from viral sequence data. *Molecular Biology and Evolution* 29: 347–357.

- [169] Drummond A, Rambaut A (2007) BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evolutionary Biology* 7: 214.
- [170] Bouckaert R, Kuhnert D, Vaughan TG, Wu CH, Xie D, et al. (2013). BEAST2: A software platform for Bayesian evolutionary analysis. available at <http://beast2.cs.auckland.ac.nz>.
- [171] Parker J, Rambaut A, Pybus OG (2008) Correlating viral phenotypes with phylogeny: Accounting for phylogenetic uncertainty. *Infection, Genetics and Evolution* 8: 239–246.
- [172] Morens DM, Taubenberger JK (2010) Historical thoughts on influenza viral ecosystems, or behold a pale horse, dead dogs, failing fowl, and sick swine. *Influenza and Other Respiratory Viruses* 4: 327–337.
- [173] Glass K, Wood JLN, Mumford JA, Jesset D, Grenfell BT (2002) Modelling equine influenza 1: a stochastic model of within-yard epidemics. *Epidemiol Infect* 128: 491–502.
- [174] Mills CE, Robins JM, Lipsitch M (2004) Transmissibility of 1918 pandemic influenza. *Nature* 432: 904–906.
- [175] Wallinga J, Lipsitch M (2007) How generation intervals shape the relationship between growth rates and reproductive numbers. *Proc R Soc B* 274: 599–604.
- [176] Fraser C, Donnelly CA, Cauchemez S, Hanage WP, Van Kerkhove MD, et al. (2009) Pandemic potential of a strain of influenza A (H1N1): early findings. *Science* 324: 1557.
- [177] McMichael AJ (2004) Environmental and social influences on emerging infectious diseases: past, present and future. *Phil Trans R Soc Lond B* 359: 1049–1058.
- [178] Manolopoulou I, Matheu MP, Cahalan MD, West M, Kepler TB (2012) Bayesian Spatio-Dynamic Modeling in Cell Motility Studies: Learning Nonlinear Taxic Fields Guiding the Immune Response. *J Am Stat Assoc* 107: 855–865.
- [179] Rowan T (1990) Functional stability analysis of numerical algorithms. Ph.D. thesis, University of Texas at Austin, Austin.

- [180] King A (2008) subplex: Subplex optimization algorithm. R package version 11-3 .
- [181] Meyers LA (2007) Contact network epidemiology: Bond percolation applied to infectious disease prediction and control. *Bull Am Math Soc* 44: 63–86.
- [182] van den Driessche P, Watmough J (2002) Reproduction numbers and sub-threshold endemic equilibria for compartmental models of disease transmission. *Math Biosci* 180: 29–48.