INTERACTIONS BETWEEN SOCIAL STRUCTURE, CONTACT NETWORKS, AND INFECTIOUS DISEASE SPREAD IN WILDLIFE POPULATIONS

A Dissertation
submitted to the Faculty of the
Graduate School of Arts and Sciences
of Georgetown University
in partial fulfillment of the requirements for the
degree of
Doctor of Philosophy
in Biology

By

Pratha Sah, M.S.

Washington, DC
October 16, 2017
Socially complex species that live in large groups are traditionally considered to have elevated risks of disease transmission. Beyond group size, it is being increasingly recognized that the dynamics of infection spread largely depends on the organization of contacts occurring in host populations. Network theory has emerged as a powerful tool in recent years, as it provides a framework for incorporating host contact patterns into models of infectious disease spread. Comparing empirical networks has proven to be challenging, however, and limited knowledge about disease transmission mechanisms has resulted in underutilization of network analytic approaches in wildlife epidemiology.

In this dissertation, I address these challenges to answer three specific questions - (i) does social structure mitigate the risk of disease transmission?, (ii) can unique social network structures associated with different social systems predict their disease outcomes?, and (iii) can empirical networks be utilized to infer transmission pathways of infectious diseases? I answer these questions by leveraging an extensive empirical dataset of the social organization of more than 40 species, and developing novel tools for network analysis.

I first test the hypothesis that subdivisions in animal social networks alleviate the disease costs of group living. Since direct observations of social interactions in solitary species and species with cryptic lifestyle are difficult, I first demonstrate how bipartite
networks of refugium use can be used to understand the hidden social structure of such species. My analysis of empirical and theoretical networks reveals that infectious disease spread is largely unaffected by the underlying modular subdivisions except when host populations are extremely subdivided. Next, I perform a meta-analysis of more than 600 animal social networks to investigate the disease implications of species’ social system. I find that only few features of social structure distinguish different social systems, and that the network organization in social species may not provide general protection against pathogens of various transmission potential. Finally, because transmission pathways of many wildlife infectious diseases tend to be unknown, I develop a tool that estimates the statistical power of empirical contact networks to predict infectious disease spread, and enables hypothesis testing between different network models.

Together, the results in this dissertation offer new perspectives on the debate about the disease costs of social living, and form a framework of integrating conventional wildlife survey techniques into network modeling. The tools developed in this dissertation may also prove useful in formulating disease prevention and conservation strategies in wildlife populations.

INDEX WORDS: animal social network, bipartite networks, community structure, contact network, meta-analysis, modularity, random graphs, sociality, wildlife disease
DEDICATION

I dedicate this dissertation to my dad, Mr. Vijay Vardhan Sah, for his unconditional love, support, and encouragement to follow my dreams.
ACKNOWLEDGMENTS

I would never be able to thank my adviser, Dr. Shweta Bansal, enough for being an incredible mentor over the past five years. From our very first exchange of emails, she has guided me in the right direction. Over the years, she has not only helped me to become a better scientist, but also a good communicator, collaborator, and colleague. In particular, I thank her for giving me the independence to explore my research ideas and learn new technical skills. I deeply admire her work ethics and dedication towards her students, and hope to become as good as a mentor as her someday. I feel honored and lucky to be her first PhD student, and hope that we continue to work together to create exciting new directions in the field.

I am grateful to my dissertation committee members, Dr. Janet Mann, Dr. Kenneth Nussear and Dr. Sivan Leviyang for their feedback and guidance, which greatly helped shape my research direction. I would especially like to thank Dr. Janet Mann for her guidance and advice on the broader picture of my dissertation.

I thank the Bansal group, for all the interesting discussions and for providing an enjoyable environment to learn. I particularly feel lucky to have shared my time at the lab with Elizabeth Lee, Dr. Ewan Colman, Dr. Eamon O’Dea, Dr. Stephan Leu and Dr. Ian Carroll, who continue to inspire me with their enthusiasm towards science, and have been brilliant role models as early career scientists.

I have been extremely fortunate to have had the opportunity to work with researchers across the world. I am especially indebted to Dr. Peter Hudson for his advice and guidance throughout my time as a PhD student. I also thank Christina
Aiello and Dr. Todd Esque for their hard work in collecting field data on desert tortoises and kindly sharing their data with me.

Finally, I would like to thank my family for being a constant support and encouragement throughout my journey. I am especially thankful to my husband, Felix Francis, for his support and patience throughout these five years and beyond.
# Table of Contents

1 Introduction ................................................................. 1

2 Inferring social structure and its drivers from refuge use in the desert tortoise, a relatively solitary species ........................................... 7
   2.1 Introduction .......................................................... 7
   2.2 Materials and methods ............................................. 11
   2.3 Results ................................................................. 19
   2.4 Discussion ............................................................ 26
   2.5 Conclusions ........................................................... 30

3 Exploring community structure in biological networks with random graphs 32
   3.1 Introduction .......................................................... 32
   3.2 Materials and methods ............................................. 37
   3.3 Results and discussion ............................................. 45
   3.4 Conclusions ........................................................... 60
   3.5 Availability and requirements .................................... 62

4 Unraveling the disease consequences and mechanisms of modular structure in animal social networks ........................................... 63
   4.1 Introduction .......................................................... 63
   4.2 Materials and methods ............................................. 64
   4.3 Results and discussion ............................................. 68
   4.4 Conclusions ........................................................... 79

5 Disease implications of animal social network structure: a synthesis across social systems .................................................. 84
   5.1 Introduction .......................................................... 84
   5.2 Materials and methods ............................................. 87
   5.3 Results ................................................................. 97
   5.4 Discussion ............................................................ 101
   5.5 Challenges and opportunities .................................... 107

6 Revealing transmission mechanisms of infectious diseases through empirical contact networks ........................................... 109
   6.1 Introduction .......................................................... 109
   6.2 Results ................................................................. 112
6.3 Discussion ......................................................... 123
6.4 Materials and methods ............................................. 128

7 Conclusions .......................................................... 134

APPENDIX

A Chapter 2 Supplementary materials .................................. 137

B Chapter 3 Supplementary materials .................................. 146
  B.1 Expected modularity .................................................. 146
  B.2 Tolerance on average-degree and average within-degree of individual modules ................................. 147
  B.3 Within-module degree distribution follows the total degree distribution .......................................... 148
  B.4 Rejection rate of degree and within-degree sequence ........... 149
  B.5 Generating disassortative modular random graphs ............. 150
  B.6 Comparing structural properties of modular random and SBM graphs ................................................. 152
  B.7 Effect of network size on network properties of modular random graphs ........................................... 153
  B.8 Effect of average network degree on network properties of modular random graphs ............................ 155
  B.9 Effect of average community size on network properties of modular random graphs ............................ 156
  B.10 Performance of other community detection algorithms on modular random graphs ............................ 157
  B.11 Accuracy of network partitioning by Lovain and fast modularity algorithm ............................................ 160
  B.12 Null analysis of empirical networks ............................... 161

C Chapter 4 Supplementary materials .................................. 164
  C.1 Animal social networks ............................................... 164
  C.2 Measure of relative modularity ....................................... 164
  C.3 Mechanisms of modular organization ............................... 166
  C.4 Generation of null networks .......................................... 167
  C.5 Generation of synthetic modular networks ....................... 167
  C.6 Disease simulations ................................................... 168
  C.7 Threshold analysis .................................................... 169

D Chapter 5 Supplementary materials .................................. 187

E Chapter 6 Supplementary materials .................................. 201

Bibliography .............................................................. 206
Figure 2.1 Critical habitat range of the desert tortoise within the Mojave desert, USA as determined by the US Fish and Wildlife Services in 2010 (http://www.fws.gov/) .................................................. 10
Figure 2.2 Bipartite network and single-mode projection of burrow use patterns at MC site during the year 2012 .................................................. 13
Figure 2.3 Frequency distribution of (a) tortoise degree i.e., unique burrows used by desert tortoises and (b) burrow degree i.e., unique tortoises visiting burrows during active (Mar-Oct) and inactive (Nov-Feb) seasons, excluding zero degree nodes .................................................. 19
Figure 2.4 Spatial constraints on asynchronous burrow associations during active seasons at study sites with control animals ................................. 21
Figure 2.5 The effect of various predictors on the two models of burrow use patterns in desert tortoises .................................................. 24
Figure 3.1 Schematic representation of the steps of random modular network generator .................................................. 41
Figure 3.2 Modular random graphs with $n=150$, $m=375$, $K=3$, $P(s=50)=1$ and $p_k$ is power law with modularity values of : a) $Q=0.1$; b) $Q=0.3$; and c) $Q=0.6$ .................................................. 47
Figure 3.3 Values of (a) assortativity, $r$, (b) clustering coefficient, $C$, and (c) path length, $L$ in modular random graphs do not vary significantly with increasing modularity ($Q$) .................................................. 49
Figure 3.4 Performance of the Louvain method, fast modularity method, the spin-glass based method, the infoMAP method, label propagation and the random-walk based method in networks with mean degree 10 .................................................. 52
Figure 3.5 Visualization of empirical and random graphs of social interaction of dolphins and food-web trophic interactions at the Little Rock Lake in Wisconsin .................................................. 55
Figure 3.6 Comparisons of empirical networks, modular random graphs and random graphs with matched degree distribution (based on the configuration model) .................................................. 56
Figure 4.1 Schematic representation of a (a) homogeneous social network with no modular subdivision, and (b-d) highly modular social networks with (b) high cohesion within subgroups, (c) high network fragmentation, and (d) high subgroup size variation ................................. 65
Figure B.3 Generating disassortative modular random graphs: Modular random graphs with \( n = 150, \ m = 375, \ K = 3, \ P(s=50) = 1 \) and \( p_k \) is power law with modularity values of: a) \( Q = -0.2 \) and b) \( Q = 0.2 \).

Figure B.4 Comparing modular random graphs and stochastic block model (SBM) graphs: Network property of (a) degree assortativity, (b) clustering coefficient and (c) path length in modular random graphs (MRG) and stochastic block model (SBM) graphs of 2000 nodes, mean degree 10 and 10 modules on increasing modularity (\( Q \)).

Figure B.5 Network property of (a) degree assortativity, (b) clustering coefficient and (c) path length in random modular graphs of mean degree 5 and 10 modules on increasing network size (\( n \)).

Figure B.6 Network property of (a) degree assortativity, (b) clustering coefficient and (c) path length in random modular graphs with 10 modules over a range of mean network degree.

Figure B.7 Network property of (a) degree assortativity, (b) clustering coefficient and (c) path length in random modular graphs of size 1000 with mean degree 10 but different number of modules.

Figure B.8 Performance of various community detection algorithms on random modular networks with Poisson degree distribution.

Figure B.9 Performance of various community detection algorithms on random modular networks with geometric degree distribution.

Figure B.10 Performance of various community detection algorithms on random modular networks with power-law degree distribution.

Figure B.11 Accuracy of partitions detected by Louvain and fast modularity algorithm in networks with mean degree 10 measured by Jaccard similarity and variation of information index.

Figure C.1 Empirical animal social networks and the corresponding homogeneous null networks included in Figure 2c.

Figure C.2 Threshold value of relative modularity above which the empirical networks tend to experience lower outbreak sizes as compared to homogeneous null networks in Figure 2c.

Figure C.3 Extended figure of Figure 2c: comparisons between real (filled points) and their homogeneous null networks (tip of arrows) with respect to the percentage of infected individuals (outbreak size) due to an outbreak with basic reproduction number, \( R_0 \) equal to 1.45, 1.8, 2.4 and 4.8.

Figure C.4 The values of network fragmentation and subgroup cohesion present in empirical animal social networks.

Figure C.5 Reduction in outbreak size experienced by empirical animal social networks as compared to their homogeneous null networks (as shown in Figure 2c) as a function of subgroup cohesion and (a) network fragmentation, and (b) subgroup size variation.
Figure C.6 Disease implications of modular subdivisions in synthetic modular networks as a function of subgroup cohesion and network fragmentation ................................................................. 176
Figure C.7 (a) Outbreak size and (b) outbreak duration in networks with different heterogeneity in subgroup sizes ................................................................. 177
Figure C.8 Disease burden in modular social networks with Poisson (purple bars) and exponential (orange bars) contact heterogeneity ................. 178
Figure C.9 Percentage error in outbreak size predictions using modular and homogeneous null networks for 19 animal social networks due to an outbreak with $R_0 = 1.05, 1.45, 1.8$ and $2.0$ ......................... 179
Figure C.10 Outbreak size predictions from modular and homogeneous null networks as compared to the "true" outbreak size (empirical networks, gray vertical bars) .................................................. 180
Figure C.11 Highly modular networks are able to reduce outbreak size: empirical evidence from the spread of Babesia parasite in the field voles (Microtus agrestis) ......................................................... 181
Figure C.12 Low modular networks do not limit global disease transmission: empirical evidence from the spread of a gastrointestinal parasite in brown spider monkeys (Ateles hybridus) ........................................ 182
Figure C.13 Structural trapping of infectious disease increases with network fragmentation: empirical evidence from the spread of pneumonia in the big horn lambs (Ovis canadensis) ........................................ 183
Figure C.14 High subgroup cohesion can induce structural delay of infection spread: empirical evidence from the spread of Mycoplasma gallisepticum in house finches (Haemorhous mexicanus) .................. 184
Figure C.15 Modular networks influence disease transmission of moderately, but not low, contagious pathogens/parasites ................................. 185
Figure D.1 Average proportion of weak ties of individuals across the three social systems ................................................................. 187
Figure D.2 Role of network structures in influencing disease transmission .................................................. 189
Figure D.3 Disease costs of social organization due to interaction network structure .................................................. 190
Figure D.4 Disease implications of social systems due to pathogens with increasing transmissibility .................................................. 191
Figure D.5 Role of network structures in influencing transmission of disease with an infectious period of 10 days ($\gamma =0.1$) .................................................. 192
Figure D.6 Disease costs of social systems due to social network structure .................................................. 193
Figure D.7 Disease costs of the three social systems when ties less than 20% of highest edge weight are removed from the social networks .................................................. 194
Figure E.1 Robustness plot of (a) INoDS, (b) $k$-test and (c) network position test to three common forms of missing data - missing data, missing edges and missing infection cases .................................................. 201
Figure E.2  Estimation of the social ($\beta$) and the asocial ($\alpha$) transmission parameter by INoDS under three forms of missing data conditions - (a) missing nodes, (b) missing edges and (c) missing infection cases. 202

Figure E.3  Relative error in the estimations of parameter $\beta$ under missing data conditions with and without the inclusion of the asocial parameter ($\alpha$) in the INoDS formulation. 202

Figure E.4  Relationship between social and asocial force of infection with increasing percentage of missing data. 203

Figure E.5  Plot of sensitivity and specificity of (a) INoDS, (a) k-test and (c) network position test to three common forms of missing data - missing nodes, missing edges and missing infection cases. 204

Figure E.6  Identifying the contact network model of $Crithidia$ spread in bumble bee colony (colony UN1). 205
List of Tables

Table 2.1  Model variables considered to characterize burrow use patterns in the desert tortoise, *Gopherus agassizii* .......................... 16
Table 4.1  Multivariable analysis on determinants of modular organization \((Q_{rel})\) across groups of different species. ................................. 71
Table 5.1  Effect size estimates of the Bayesian generalized linear mixed models examining the characteristics of social network structure among the three social systems: relatively solitary, gregarious and socially hierarchical ................................................. 96
Table A.1  Location, surveyed years, total tortoises tracked and average observations at the study sites ........................................ 137
Table A.2  Location of the weather stations used to obtain climatic information for the nine study sites ........................................ 142
Table A.3  Network density, degree homophily, clustering coefficient, modularity and degree centralization of tortoise social networks at five study sites with control animals .................................................. 143
Table A.4  Correlation between geographical distances and edge occurrence in tortoise social networks during active season of each surveyed year 144
Table A.5  Potential interactions considered for the burrow switching and the burrow popularity model ........................................ 145
Table B.1  Comparisons of random modular networks with different degree- and within-degree distributions type ................................ 162
Table B.2  Comparisons of empirical and random networks with randomized within-edge and between-edge connections ................ 163
Table C.1  Multivariable analysis on determinants of modular organization in the dynamic networks of four species (*C. fellah, C. pennsylvanicus, P. lotor*, and *M. agrestis*) ................................................. 170
Table C.2  Empirical (or data-driven modeling) evidence that matches the predictions of our study .................................................. 186
Table D.1  Description of the network measures used in this study ...... 188
Table D.2  Percentage of variance explained by the random effects included in the Bayesian generalized linear mixed model presented in Table 1 (assuming relative solitary species as the base social system) ...... 195
Table D.3  Results of cross validation using random sub-sampling method 195
Table D.4 Effect size estimates of two separate phylogenetically controlled Bayesian generalized liner mixed models examining the characteristics of social networks associated with (a) the type of interaction recorded and (b) the spatial scale of data collection

Table D.5 Filtered networks with 5% weakest edges removed: Effect size estimates of the generalized linear mixed models (by MCMCglmm) examining the characteristics of filtered interaction network structure for different species sociality

Table D.6 Filtered networks with 10% weakest edges removed: Effect size estimates of the generalized linear mixed models (by MCMCglmm) examining the characteristics of filtered interaction network structure for different species sociality

Table D.7 Filtered networks with 15% weakest edges removed: Effect size estimates of the generalized linear mixed models (by MCMCglmm) examining the characteristics of filtered interaction network structure for different species sociality

Table D.8 Filtered networks with 20% weakest edges removed: Effect size estimates of the generalized linear mixed models (by MCMCglmm) examining the characteristics of filtered interaction network structure for different species sociality
Chapter 1

Introduction

Infectious disease spread is primarily driven by the organization of infection spreading interactions among individuals. Disease predictions from traditional epidemiological models such as compartmental [1], metapopulation [2], and lattice-based models [3] can be unreliable as they ignore heterogeneity in individual behavior and assume homogeneous mixing of individuals in a population. In recent years, network analytic approaches have gained momentum as they capture the variation in individual behavior as well as the local, and global patterns of contacts present in the host populations [4, 5, 6, 7]. Formally, networks are mathematical representations of interactions among components of a system. In network epidemiology, contact networks typically describe a set of individuals (represented as nodes) connected by the occurrence of infection-spreading contacts (represented as edges). Infectious disease spread can also be modeled at a higher spatial scale by representing households, airports or populations in a metapopulation as nodes. Edges in these models represent movement of individuals between nodes.

Network epidemiology in the past has been successfully used to predict transmission of infectious human diseases such as HIV [8], SARS [4], and livestock diseases such as the foot-and mouth epidemic in the UK [9]. Networks have also recently gained popularity in the field of wildlife epidemiology [10, 11, 12, 13]. A straightforward approach to investigate the disease risks of a wildlife population is to first construct a contact network by empirically collecting data on each disease spreading interaction.
occurring in the population, and then performing numerical or analytical investigations of the disease outcomes. However, given the logistical difficulties of observing all individuals and interactions occurring in a population, data on wild animal populations is usually limited precluding the construction of complete contact networks. In such cases, observations of coincident use of spatial resources can be used to construct "proxy" networks of social structure. Such a technique is particularly useful for solitary species or species with cryptic lifestyles, where direct observations of social interactions are difficult, and therefore integration of behavior in conservation and management has been limited. For such species, quantifying spatial behavior can provide insights towards hidden social structure, establish proxy contact patterns relevant for infectious disease spread, and also provide early warning signals of the presence of population stressors that may pose a threat to the health of the populations.

In Chapter 2, we tackle the problem of inferring population-level social patterns of a relatively solitary species, the desert tortoise. Specifically, we investigate the role of various demographic and environmental factors that influence burrow use behavior of desert tortoises. In addition, we identify the factors that driver the popularity of burrows in the desert tortoise habitat. We quantify burrow use behavior using bipartite networks, which are networks with two types of nodes - animals and burrows. Burrow use by tortoises are indicated by placing edges connecting animal and burrow nodes. Since management decisions primarily focus on the consequences of threats at the population level, our approach of combining graph-theoretic and statistical methods can allow conservation biologists to understand how heterogeneity in individual behavior scales up to population level impacts. Identification of active popular refuges of wildlife species can be used to monitor population health, identify core habitat for conservation and augment routine survey methods to improve accuracy of population density estimation.
For species where direct observations of social contact are feasible but recording data on all infection spreading interactions is difficult, an alternate approach to constructing a complete contact network model is to generate synthetic "null" network models that utilize only partial information about the empirical systems. Null network models lead to a better understanding of the mechanisms behind the organization of social contacts in host populations; it also helps in predicting the dynamics of processes that traverse these networks. A basic synthetic network model utilizes the information about the total number of individuals, and the degree distribution (i.e., heterogeneity in social contacts) to generate contact network that is random with respect to other network metrics. Basic synthetic null networks are fairly easy to generate but can have a low predictive power in predicting transmission dynamics as they ignore local and global complexity of contact networks. Recently, null network models that incorporate higher-order features such as clustering [14], degree assortativity [15], and network motifs [16] have been used to examine the functional roles of such global network features in biological systems.

In Chapter 3, we develop an algorithm that generates connected random networks with a tunable strength of community structure. Community structure, or the presence of groups of nodes that are more densely connected to those within their group than others, is known to be ubiquitous in biological and social systems [17, 18, 19, 20, 21, 22, 23]. Even though the dynamics of infection spread has been studied extensively in spatially structured metapopulations [24, 25, 26, 27], there is still a lack of comprehensive understanding of infectious disease spread in heterogeneous animal populations that show subgrouping in their contact structure. This is because disentangling the effect of community structure from other network properties such as clustering coefficient and assortativity can be a challenge in empirical networks. The algorithm we introduce generates networks with a specified degree
distribution and community structure while maintaining a graph structure that is as random as possible. The generated modular networks therefore allow for the systematic study of the presence of modularity in contact networks and its impact on disease spread through host populations. In addition, the networks can serve as null models for detecting structural properties beyond the byproduct of the degree and community structure patterns and uncovering mechanistic design principles in empirical biological (and other) complex networks.

Community structure emerges in animal groups when subsets of conspecifics consistently interact with each other more often than they do with other individuals in the group, forming subgroups. Subgroup structure is common in host populations where individuals are spatially segregated [28, 29], in fission-fusion animal societies [30, 31, 32], and other group living species (e.g., sperm whales [33]). Recently, it has been proposed that the presence of subdivisions reduces the disease costs of social living [34, 35]. However, testing hypotheses about subdivisions in complex network structures of animal societies has been a challenge. This is because the popular network measure of subdivision strength, Newman modularity [36], is network-specific and cannot be used comparatively across different animal social networks [37]. In addition, it is difficult to isolate the role of community structure from other correlated higher order features of empirical social networks.

In Chapter 4, we overcome the aforementioned challenges by (i) introducing a new measure of compartmentalization, called relative modularity, that can be used to compare social networks of different animal species, (ii) generating synthetic modular networks to identify when and how modular subdivisions alleviate disease burden, and (iii) using a hierarchy of null network models to analyze the disease consequences of modular subdivisions in real animal social networks. To gain insights towards the mechanistic features that create subdivisions in animal societies, we compare empir-
ical social networks of 43 non-human species. We also perform numerical simulations of infectious disease spread on networks generated using the model developed in Chapter 3 to systematically investigate the role of community structure in mediating disease risks while controlling for other higher order features of social networks. Overall, this study provides insights towards the context and mechanisms under which community structure in animal social groups influences the spread of infectious diseases. We further suggest the use of appropriate null models that predict disease outcome in social groups with reasonable accuracy when data-limited estimates of epidemic consequences are necessary.

The variation in the structure and organization of modular subdivisions in animal social networks suggests that the evolutionary trajectory of social structure in species is shaped by the relative trade-offs of social living. Species that live in large groups and are considered to have elevated costs of disease transmission due to high contact rates [38, 39]. Beyond group size, the spread of infectious disease is largely influenced by the organization of social structure in host populations [40, 41, 42, 43]. A systematic examination of the disease costs associated with species’ social system requires a comparative approach that isolates unique structural characteristics of social connections, while controlling for population size, data collection methodology and type of interaction recorded. However, comparing networks across different taxonomic groups has proven to be a difficult task, with only a few cross-species network comparisons previously published in the literature [44, 45, 46].

In Chapter 5, we undertake a population scale investigation to understand the disease implications of species’ social structure. We use more than 650 social networks of 47 animal species to answer a specific question - how are the social networks organized across different social systems, and what are the disease implications of such differences? We overcome the challenges of comparing social networks across tax-
onomic boundaries by adopting a phylogenetically-controlled Bayesian mixed modeling approach, that accounts for differences in data collection methods, network edge weighting criteria, sampling effort, etc. To our knowledge, this is the first comparative study that utilizes an extensive social network dataset of multiple animal taxonomic groups, social systems and type of interaction recorded between animals.

In recent years, the use of modern technology such as RFID, GPS, radio tags and proximity loggers has enabled collection of movement and contact data, making network modeling feasible. Nonetheless, limited knowledge about the spreading pathogen and mode of transmission precludes an exact definition of disease causing contact which is central to constructing accurate contact network models. In Chapter 6, we address this challenge by developing a computational tool that estimates evidence supporting contact networks that represent competing hypotheses about transmission modes. Our tool accurately identifies the underlying contact network even when the networks are partially sampled and information on disease spread is incomplete. Consequently, this tool provides critical insights towards biological mechanisms of pathogen spread that can help inform early disease control strategies in cases where laboratory techniques of unraveling transmission mechanisms may take years to resolve.

From identifying the disease implications of social network structures, to providing insights about the transmission mechanisms of a novel pathogen using a network analytic approach, this dissertation aims to extend the potential of network modeling in the field of wildlife epidemiology by developing novel tools and methodological approaches that overcome the current challenges faced in the field.
2.1 Introduction

Social structure of wildlife populations is typically derived from observational studies on direct social interactions [e.g., affiliative interactions in primates [13, 34], group association in dolphins [47] and ungulates [10, 48], food sharing in vampire bats [49]]. In relatively solitary species, individuals spend a considerable amount of time alone and have minimal direct interactions with conspecifics except during mating and occasional aggressive encounters [50]. Examples of such species include raccoons, red foxes, orangutans, and some species of bees, wasps and bats. For these wildlife populations, social interactions may be limited to certain areas within their habitat, such as refuges (e.g., roost, den, burrow, nest) or watering holes that provide increased opportunities of direct contact between individuals. Monitoring these resources can therefore help establish relevant social patterns among individuals.

In addition to establishing social structure, refuges provide shelter, protection from predators and serve as sites for nesting and mating. Refuge use patterns of individuals are therefore central to survival, mating and foraging success and can

\footnote{This chapter has been published as: Sah, P., Nussear, K. E., Esque, T. C., Aiello, C. M., Hudson, P. J., and Bansal, S. (2016). Inferring social structure and its drivers from refuge use in the desert tortoise, a relatively solitary species. Behavioral ecology and sociobiology, 70(8), 1277-1289.}
serve as efficient indicators of population disturbances. Unlike traditional population
dynamics indicators such as mortality and birth rate, refuge use behavior can respond
instantaneously to sub-optimal conditions [51, 52]. Altered patterns of refuge use
may thus indicate a disturbance or change in population fitness and provide an early
warning to conservation biologists. Changes in habitat or refuge use have indeed been
linked to the presence of natural population stressors such as increased predation [53],
drought [54, 55] and disease transmission risk [56], as well as anthropogenic population
stressors of translocation [57] and urbanization [58].

While the importance of refuge use in social interactions, survival and mating
success, as well as indicators of environmental and anthropogenic stressors has been
long appreciated, biologists are only beginning to understand individual level hetero-
geneity in refuge use and its population-level consequences in relatively solitary species
[30, 59, 60]. The general absence of studies quantifying pairwise interactions due to
preferences in refuge-use implies a lack of knowledge of the baseline social organiza-
tion that could be used to evaluate changes in robustness or health of these wildlife
populations. To overcome these shortcomings we explore a modeling framework that
combines network theory with statistical models to infer the presence of and mecha-
isms behind the social organization in the desert tortoise, \textit{Gopherus agassizii}, formed
by their refuge-use preferences. The desert tortoise is a long-lived, terrestrial species
that occurs throughout the Mojave Desert north and west of the Colorado River.
Individuals of this species use subterranean burrows as an essential adaptation to
obtain protection from temperature extremes and predators. Because tortoises spend
a majority of their time either in or near burrows, most of their social interactions
are associated with burrows [61].

Social behavior in desert tortoises is not well understood, though evidence suggests
the presence of dominance hierarchies [62, 63] which may influence social structure and
burrow choice in desert tortoises. In addition to social hierarchies, previous research suggests factors such as sex [64], age [65], season [61]; and environmental conditions [66, 67] may influence burrow use in desert tortoises. If conspecific cues and environmental factors exhibit strong influence on burrow use, population stressors impacting these characteristics could alter typical burrow behavior. The two major population threats that have been identified in desert tortoise populations include upper respiratory tract disease (URTD) caused by *Mycoplasma agassizii* and *Mycoplasma testudineum* [68, 69, 70], and extreme environmental conditions, particularly drought [71, 72]. In addition to these threats, the primary management strategy in desert tortoises is to translocate animals out of areas affected by anthropogenic disturbances [73]. Translocation in other reptilian species, however, has had limited success due to high rates of mortality [74, 75] and may also act as a population stressor. In desert tortoises, all three population stressors have been linked to differences in individual behavior [66, 76, 77]. Although previous studies provide insights towards potential factors that may affect burrow use, we lack a mechanistic understanding behind the role of these factors in driving heterogeneity in burrow use patterns in desert tortoises. A large impact of population stressors on refuge use can affect mating and foraging opportunities of desert tortoises and also reduce their likelihood of survival.

In this study, we combine data-sets from nine study sites in desert tortoise habitat, spanning more than 15 years to derive burrow use patterns of individuals in these populations. We first construct bipartite networks to infer their social associations due to asynchronous use of burrows. We then use generalized linear mixed models to explain mechanisms behind heterogeneity in burrow use behavior of individuals and effect of population stressors. As the desert tortoise is a long lived species, evaluating the impact of population stressors on burrow use patterns provides an efficient alternative to using traditional demographic metrics (such as mortality). We also
Figure 2.1: Critical habitat range of the desert tortoise within the Mojave desert, USA as determined by the US Fish and Wildlife Services in 2010 (http://www.fws.gov/). Critical habitat is defined as those geographical areas that contain physical or biological features essential to the conservation and management of the species [78]. Points represent centroids of survey sites where tortoises were monitored using radio-telemetry. Point size is proportional to the number of animals monitored at the site. Site abbreviations: BSV - Bird Spring Valley, CS - Coyote Springs, FI - Fort Irwin, HW - Halfway, LM - Lake Meade, MC - McCullough Pass, PV - Piute Valley, SG - St. George, SL - Stateline Pass.
investigate the use of burrows through a bipartite network model to identify why certain burrows are more popular than others in desert tortoise habitat. Overall, our analysis of refuge-based associations provide further insights into the structure and dynamics of social organization in a species traditionally considered as solitary and provides mechanisms behind individual variation within these associations.

2.2 Materials and methods

2.2.1 Dataset

We combined datasets from nine study sites monitored from 1996 to 2014 across desert tortoise habitat in the Mojave desert of California, Nevada, and Utah (Fig. 2.1). Each site was monitored over multiple years, but not all sites were monitored in each year of the 15 year span. At each site, individuals were monitored at least weekly during their active season and at least monthly during winter months using radio telemetery. The total number of animals sampled and average number of observations per tortoise at each site is included in Supplementary Table S1. All tortoises were individually tagged, and during each tortoise encounter, data were collected to record the individual identifier, date of observation, GPS location, micro-habitat currently used by the animal (e.g., vegetation, pallet, or a burrow), any visible signs of injury or upper respiratory tract disease. As the dataset involved monitoring tagged individuals, it was not possible to record data blind. The unique burrow identification (id) was recorded for cases where an animal was located in a burrow. New burrow ids were assigned when an individual was encountered at a previously unmarked burrow.
2.2.2 Network Analysis

We constructed bipartite networks of asynchronous burrow use in desert tortoises for active (March - October) and inactive season (November - February) of each year at five sites (CS, HW, MC, PV, SL) where no translocations were carried out. An example of a burrow use bipartite network is shown in Fig. 2.2. The network consisted of burrow and tortoise nodes and undirected edges. An edge connecting a tortoise node to a burrow node indicated burrow use by the individual (Fig. 2.2). To reduce bias due to uneven sampling, we did not assign edge weights to the bipartite networks. Edges in a bipartite network always connect the two different node types, thus edges connecting two tortoise nodes or two burrow nodes are not permitted. Tortoise nodal degree in the bipartite network therefore denotes the number of unique burrows used by the individual and burrow nodal degree is the number of unique individuals visiting the burrow. Networks were generated using Networkx package in Python [79].

We further examined the social structure of desert tortoises by converting the bipartite network into a single-mode projection of tortoise nodes (Tortoise social network, Fig. 2.2). For these tortoise social networks, we calculated network density, degree centralization, modularity, clustering coefficient, and homophily of individuals by degree and sex/age class. Network density is calculated as the ratio of observed edges to the total possible edges in a network [50]. Degree centralization measures the variation in node degree across the network, such that high values indicate a higher heterogeneity in node degree and that a small proportion of nodes have a higher degree than the rest [50]. Modularity measures the strength of the division of nodes into sub-groups [80] and clustering coefficient measures the tendency of neighbours of a node to be connected [14]. The values of modularity and clustering coefficient can range from 0 to 1, and larger values indicate stronger modularity or clustering coefficient.
Figure 2.2: (a) Bipartite network of burrow use patterns at MC site during the year 2012. Node type indicated by color (Blue = adult males and red = adult females). Node positions are fixed using Yifan Hu’s multilevel layout in Gephi [81]. In this paper, we quantify burrow switching and burrow popularity as degree of tortoise nodes and burrow nodes, respectively, in the bipartite network. For example, burrow switching of the female tortoise X is five and burrow popularity of burrow Y is one. (b) Single-mode projection of the bipartite network into tortoise social network. Nodes with zero degree have been removed for clarity of illustration.

We generated 1000 random network counterparts to each empirical network using double-edge swap operation in NetworkX [79] to determine if the observed network metrics were significantly different from random expectation. The generated random networks had the same degree sequence as empirical networks, but were random with respect to other network properties.

We next examined the spatial dependence of asynchronous burrow associations by using coordinates of burrows visited by tortoises to calculate centroid location of each tortoise during a particular season of a year. Distances between each tortoise pair \((i, j)\) were then calculated as \(d_{ij} = d_{ji} = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}\) where \((x, y)\) is the coordinate of tortoise centroid location. Pearson correlation coefficient was used to calculate the correlation between observed edges in social network and geograph-
ical distances between the tortoises. We compared the observed correlation to a null
distribution of correlation values generated by randomly permuting spatial location
of burrows 10,000 times and recalculating correlation between social associations and
distance matrix for each permutation. Correlation were calculated using MantelTest
package in Python [82].

2.2.3 Regression Analysis

We used generalized linear mixed regression models with Poisson distribution and log
link function to assess burrow use patterns. To capture seasonal variation in burrow
use, we aggregated the response counts over six periods (Jan-Feb, Mar-Apr, May-
Jun, Jul-Aug, Sep-Oct and Nov-Dec). Patterns of burrow use were analyzed in two
ways. First, we investigated factors affecting burrow switching, which we define as
the number of unique burrows used by a tortoise in a particular sampling period.
Second, we investigated burrow popularity, defined as the number of unique individ-
uals using a burrow in a particular sampling period. Model variables used for each
analysis are summarized in Table 2.1. All continuous model variables were centered
(by subtracting their averages) and scaled to unit variances (by dividing by their stan-
dard deviation). This standard approach in multivariate regression modeling assigns
each continuous predictor with the same prior importance in the analysis [83]. All
analyses were performed in R (version 3.0.2; R Development Core Team 2013).

Investigating burrow switching of desert tortoises

In this model, the response variable was burrow switching, defined as the total number
of unique burrows used by desert tortoises during each sampling period. An individual
was considered to be using a burrow if it was reported either inside a burrow or within
25 m² grid around a burrow. The predictors included in the model are described in
Table 2.1. In addition to the fixed effects, we considered three interactions in this model (i) sampling period × sex, (ii) sampling period × seasonal rainfall and (iii) local tortoise density × local burrow density. Tortoise identification and year × site were treated as random effects.

INVESTIGATING BURROW POPULARITY

For this model, the response variable was burrow popularity defined as the total number of unique tortoises using a focal burrow in a sampling period. The predictors included in the model are also described in Table 2.1. In this model, we also tested for three interactions between predictors including (i) sampling period × seasonal rainfall, (ii) sampling period × local tortoise density, and (iii) local tortoise density × local burrow density. We treated burrow identification and year × site as random effects.

POPULATION STRESSORS

Disease: We considered tortoises exhibiting typical signs of URTD including nasal discharge, swollen (or irritated/sunken) eyes, and occluded nares to be indicative of an unhealthy animal. As diagnostic testing was not the focus of the studies collecting the data, we were unable to confirm the infection status of individuals. Knowledge of confirmed infection status of animals, however, was not central to our study as our aim was to measure behavioral response of symptomatic individuals only. We included health condition in the regression model as a categorical variable with two levels - healthy and unhealthy. An individual was considered to be unhealthy if it was reported to display clinical signs of URTD at least once during the sampling period.

Translocation: We accounted for translocation in the regression model by giving each surveyed tortoise one of the following five residency status at each sampling period:
Table 2.1: Model variables considered to characterize burrow use patterns in the desert tortoise, *Gopherus agassizii*.

<table>
<thead>
<tr>
<th>Variables</th>
<th>Variable type</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Tortoise attributes (Burrow switching model only)</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sex/age class</td>
<td>Categorical</td>
<td>Three levels - adult males, adult females and non-reproductives (including neonates, juveniles and subadults)</td>
</tr>
<tr>
<td>Size</td>
<td>Continuous</td>
<td>Midline carapace length averaged over the year for each individual</td>
</tr>
<tr>
<td><strong>Burrow attributes (Burrow popularity model only)</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Burrow azimuth</td>
<td>Categorical</td>
<td>Direction in which burrow entrance faces forward. We converted the 1 to 360° range of possible azimuth values to eight categorical azimuth directions: Q1 (1-45), Q2 (46-90), Q3 (91-135), Q4 (136-180), Q5 (181-225), Q6 (226-270), Q7 (271-315) and Q8 (316-360)</td>
</tr>
<tr>
<td>Burrow surveyed age</td>
<td>Continuous</td>
<td>Number of years between the first report of burrow and current observation</td>
</tr>
<tr>
<td>Soil condition</td>
<td>Categorical</td>
<td>The soil conditions at the nine sites varied from sandy to mostly rocky. We therefore categorized burrow soil into four categories - mostly sandy, sand and rocky, mostly rocky and caliche and rocky</td>
</tr>
<tr>
<td>Percentage wash</td>
<td>Continuous</td>
<td>Percentage area covered by dry bed stream within 250 sqm area around burrow</td>
</tr>
<tr>
<td>Surface roughness</td>
<td>Continuous</td>
<td>See [84]</td>
</tr>
<tr>
<td>Topographic position</td>
<td>Continuous</td>
<td>Index of landscape elevation around 250 sqm of burrow. High values are indicative of dry lakebeds or valley bottoms, and low values represent ridges and mountain tops. See [84] for details.</td>
</tr>
<tr>
<td><strong>Environmental characteristics</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sampling period</td>
<td>Categorical</td>
<td>The period of observation as described before. We divided a year into six periods of two months each</td>
</tr>
<tr>
<td>Seasonal rainfall</td>
<td>Continuous</td>
<td>Total rainfall recorded at weather station nearest to the study site (in inches) during a particular sampling period</td>
</tr>
<tr>
<td>Temperature</td>
<td>Continuous</td>
<td>Average, maximum and minimum temperature recorded at the weather station nearest to the study site and calculated over each sampling period in our model</td>
</tr>
<tr>
<td><strong>Population stressors</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tortoise health</td>
<td>Categorical</td>
<td>Burrow switching model only. Two categories - healthy and unhealthy</td>
</tr>
<tr>
<td>Residency status</td>
<td>Categorical</td>
<td>Burrow switching model only. Each individual was assigned one the five residency status for each sampling period - Control (C), Resident (R), Translocated (T), Ex-Resident (ER) or Ex-Translocated (ET)</td>
</tr>
<tr>
<td>Drought condition</td>
<td>Continuous</td>
<td>Average rainfall from November to February used as a proxy of drought condition for the following year</td>
</tr>
<tr>
<td><strong>Density condition</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Local tortoise density</td>
<td>Continuous</td>
<td>For burrow switching model: the average number of individuals found within 10,000 sqm grid around the focal tortoise each day of sampling period when the animal was surveyed. For burrow popularity model: number of individuals found in 10,000 sqm grid around the focal burrow averaged each surveyed day of the sampling period For burrow switching model: the average number of active burrows in 10,000 sqm grid around the focal tortoise each day of the sampling period when the animal was reported. For burrow popularity model: the number of active burrows in 10,000 sqm grid around the focal burrow. A burrow was considered to be active if it was reported to be occupied at least once during the current or any previous sampling period</td>
</tr>
<tr>
<td>Local burrow density</td>
<td>Continuous</td>
<td></td>
</tr>
<tr>
<td><strong>Survey condition</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sampling days</td>
<td>Continuous</td>
<td>Total survey days during the sampling period</td>
</tr>
<tr>
<td>Individual level bias</td>
<td>Continuous</td>
<td>Burrow switching model: Total number of days when the focal tortoise was reported using any burrow to account for any survey biases between individuals. Burrow popularity model: Total tortoises surveyed during the sampling period</td>
</tr>
</tbody>
</table>
Control (C), Resident (R), Translocated (T), Ex-resident (ER) or Ex-translocated (ET). Translocations were carried out at four (BSV, FI, LM, SG) out of nine sites in our dataset for purposes described in previous studies [76, 85]. All animals native to the site were categorized as Controls (C) during sampling periods before translocation occurred. For sampling periods post translocation, all native animals were categorized as Residents (R), and introduced animals were categorized as Translocated (T). One year after translocation, translocated and resident tortoises were considered to be Ex-translocated (ET) and Ex-residents (ER), respectively, to account for potential acclimatization of introduced animals [76]. We note that one of the four translocation sites (SG) did not have native animals prior to translocation. No translocations were carried out at the rest of the five sites, so all animals surveyed at those sites were labeled as controls in all sampling periods.

**Drought:** The desert tortoise habitat in Mojave desert typically receives most of the rainfall during the winter season. We therefore used winter rainfall to assess drought conditions in desert tortoise habitat. We defined winter rain during a year as average rainfall from November to February and used it as a proxy of drought condition for the following year. We note that summer rainfall in desert tortoise habitat varies from west to east, where summer rainfall becomes a larger component of the total annual precipitation in East Mojave desert [86]. Therefore, although we used winter rainfall as a proxy of drought conditions, we considered the effects of summer precipitation implicitly by including seasonal rainfall as a separate predictor (see Table2.1).

**Model selection and validation**

Following [87] we avoided model selection to remove non-significant predictors and instead present results of our full model. Using the full model allows model predictions conditional on the values of all predictors and results in more accurate confidence
interval of effects of interest [87]. The Bayesian information criterion (BIC) of model selection was used to identify the best higher order interactions. A potential drawback of including all independent variables in the final model is multicollinearity. We therefore estimated Generalized Variance Inflation Factor (GVIF) values for each predictor. GVIF is a variant of traditional VIF used when any predictor in the model has more than 1 degree of freedom [88]. To make GVIF comparable across dimensions, [88] suggest using \( \text{GVIF}^{(1/(2Df))} \) which we refer to as adjusted GVIF. We sequentially removed predictors with high adjusted GVIFs, recalculated adjusted GVIF, and repeated the process until all adjusted GVIF values in the model were below 3 [89].

We carried out graphical diagnostics by inspecting the Pearson residuals for the conditional distribution to check if the models fit our data in each case. We detected under-dispersion in both the regression models. Under-dispersed models yield consistent estimates, but as equidispersion assumption is not true, the maximum-likelihood variance matrix overestimates the true variance matrix which leads to over-estimation of true standard errors [90]. We therefore estimated 95\% confidence intervals of fixed and random effects using bootstrapping procedures implemented in 'bootMER' function in package lme4.

We tested for the significance of fixed factors in both the models using likelihood ratio test (R function mixed from afex package [91]). For significant categorical predictors, we used Tukeys HSD (R function glht from the multcomp package, [92]) as a post-hoc test of significant pair-wise differences among means. All reported P-values of post-hoc tests are adjusted for multiple comparisons using the single-step method [92].
2.3 Results

2.3.1 Network Analysis

Bipartite networks of asynchronous burrow use across all sites demonstrated considerable variation in degree of tortoise nodes and burrow nodes (Fig. 2.3). Tortoises visited more unique burrows on average (4.03 ± 3.43 SD) and had a greater range of burrows visited in active seasons (1-9) than in inactive seasons (average = 1.46 ± 0.72 SD, range = 1-5). Less than 40% of tortoises used more than one burrow during Nov-Feb (inactive) months (Fig. 2.3a). Most burrows in desert tortoise habitat were visited by a single tortoise during active and inactive season (Fig. 2.3b). Heterogeneity in the number of animals visiting burrows, however, tended to be slightly more during the months of March-November than November-February (active = 1.21 ± 0.56 SD, inactive = 1.08 ± 0.35 SD).
The tortoise social network (constructed as a single mode projection of tortoise nodes from the bipartite network) demonstrated moderate clustering coefficient (0.36 ± 0.21 SD) and modularity (0.53 ± 0.15 SD). Twenty three of the 24 social networks we analyzed had higher clustering coefficient and 18 social networks were more modular than random networks (Supplementary Table S3). Thirteen social networks out of the total 24 demonstrated significant degree homophily (when nodes with similar degree tend to be connected) and 11 of those had positive associations (Supplementary Table S3). Positive degree homophily suggests that tortoises using many unique burrows often use the same set of burrows and are therefore connected in the social network. Tortoise social networks also had a moderate positive degree centralization which indicates a small subset of individuals used more burrows than the rest in the sampled population. Within sexes, positive degree centralization was observed both within males (0.20 ± 0.08 SD) and females (0.17 ± 0.06 SD). Homophilic association by sex ranged from -0.6 to 0.11 indicating a preference for one sex to associate with the opposite. These negative sexwise associations, however, were not different than those expected by chance.

The association between tortoises in their social network was inversely correlated with geographical distances between them, indicating that individuals closer to each other preferred using the same set of burrows. The magnitude of correlation ranged from -0.22 to -0.89 with an average value of -0.49 (Fig. 2.4). The P-value of the permutation test for all sites across active seasons of all surveyed years was less than 0.05, indicating a significant effect of geographical location on social associations (Supplementary Table S4). This result of spatial constraints driving social interactions is not surprising as geographical span of surveyed sites were much larger (>1500m) than the normal movement range of desert tortoises [67]. However, the moderate
Figure 2.4: Spatial constraints on asynchronous burrow associations during active seasons at study sites with control animals. At each site, correlation is calculated between geographical distance and edge occurrence in tortoise social network, and averaged over each surveyed year. Error bars represent standard errors with n=8 (CS), n=3 (HW), n=2 (MC), n=7 (PV) and n=2 (SL). P-value associated with each correlation measure is < 0.05.

The value of correlations suggest other factors (such as environmental, social, density) could play an important role in desert tortoise’s asynchronous burrow associations.

2.3.2 Regression Analysis

Based on the observed heterogeneity in bipartite networks, we next investigated the relative effect of natural variables and population stressors on burrow switching patterns of desert tortoises (viz degree of animal nodes in bipartite networks) and popularity of burrows in desert tortoise habitat (viz degree of burrow nodes in bipartite networks). Supplementary Table S5 presents the best models of BIC values for interactive predictors that explain burrow switching in desert tortoises and burrow popularity. The three interactions tested for burrow switching models were sampling
period × sex, sampling period × seasonal rainfall and local tortoise density × local burrow density. We tested all possible combinations of the three interactions. The best model contained an interaction of sampling period × seasonal rainfall (Supplementary Table S5). For the burrow popularity model, we tested all possible combinations of the sampling period × seasonal rainfall, the sampling period × the local tortoise density and the local tortoise density × local burrow density interactions. The best model included the sampling period × the local tortoise density and the local tortoise density × the local burrow density interaction term.

Multicollinearity tests revealed all three measures of temperature (average, max and min) to have adjusted GVIF values of >3. The three predictors were therefore dropped from both the models. We also removed the sampling period × tortoise density interaction from the burrow popularity model as it inflated adj GVIF value of tortoise density to >3. $\sigma^2$ estimate of tortoise identification and burrow identification random effect was negligible (tortoise identification: $\sigma^2 = 0$, CI = 0-0.004, burrow identification: $\sigma^2 = 0$, CI = 0-0.01). Both random effects were therefore removed from the regression models.

**Effect of animal attributes**

Sex/age class had a significant effect on burrow switching ($\chi^2_{2} = 16.75$, $P = 0.0002$). Overall, adults used more unique burrows than non-reproductives. Among adults, males used a slightly higher number of unique burrows than females (Fig. 2.5). There was no effect of body size on individuals’ burrow switching behavior ($\chi^2_{1} = 0.2$, $P = 0.65$).
Effect of burrow attributes

Out of the six burrow attributes included in the model, burrow age and surface roughness around burrow had the highest impact on burrow popularity, i.e., number of unique individuals visiting the burrow (burrow age: $\chi^2_1 = 46.07, P < 0.0001$, surface roughness: $\chi^2_1 = 14.37, P = 0.0002$). Burrow popularity was positively correlated with surface roughness indicating that burrows in flat sandy areas were visited by fewer unique tortoises than burrows in rough rocky areas (Fig. 2.5). Older burrows were visited by more unique individuals, with burrow popularity increasing $e^{0.08}$ times with each increment of age (Fig. 2.5). Burrows in areas with higher topographical position as indicated by GIS raster images were also more popular ($\chi^2_1 = 5.71, P = 0.02$).

Effect of environmental conditions

Sampling period had a large effect on number of unique burrows used by desert tortoises ($\chi^2_5 = 160.96, P < 0.0001$) as well as on burrow popularity ($\chi^2_5 = 176.25, P < 0.0001$) as compared to other model predictors. Burrow switching of desert tortoises was highest during the months of May-June and September-October when they are typically more active, and lowest in winter months (Fig. 2.5). In the late summer (July-August), tortoises demonstrated slightly lower burrow switching than during the active season, but higher than the winter season. Within a particular year, the direction of the effect of seasonal rainfall varied across different sampling periods (sampling period $\times$ seasonal rain: $\chi^2_5 = 107.46, P < 0.0001$). For example, high rainfall during the months of March-April reduced burrow switching in desert tortoises. On the other hand, individuals exhibited higher burrow switching with higher rain during the months of July-August (Supplementary Fig. S3b).
Figure 2.5: The effect of various predictors on the two models of burrow use patterns in desert tortoises. Error bars indicate 95% confidence intervals around the estimated coefficient value. For continuous predictors, the vertical dashed line indicates no effect - positive coefficients indicate increase in burrow popularity/switching with increase in predictor value; negative coefficients indicate decrease in burrow popularity/switching with higher values of predictors. For each categorical predictor, the base factor (solid data points) straddles the vertical line at 0 and appears without a 95% CI. Positive and negative coefficients for categorical predictors denote increase and decrease, respectively, in burrow popularity/switching relative to the base factor.
In contrast to the large variation in individuals’ burrow switching behavior between sampling periods, popularity of burrows did not vary during a large portion of the year (May - December). Total unique animals visiting burrows tended to be lower in the months of January-February and March-April, as compared to other months of the year (Fig. 2.5, S4c). Seasonal rainfall had a positive correlation with burrow popularity ($\chi^2 = 6.02, P = 0.01$).

**Effect of density conditions**

An increase in the number of active burrows around individuals promoted burrow switching, whereas an individual used fewer burrows when there were more tortoises in the vicinity (Fig. 2.5). In the burrow popularity model, higher tortoise density around burrows increased number of individuals visiting these burrows (Fig. 2.5). There was a significant interactive effect of the two density conditions on burrow popularity ($\chi^2 = 177.37, P < 0.0001$) – increase in burrow popularity with higher tortoise density was lower when there were more burrows in the vicinity of the focal burrow (Supplementary Fig. S4d).

**Effect of population stressors**

Population stressors of drought, health and translocation had variable influences on burrow switching of desert tortoises (Fig. 2.5, Supplementary Fig. 5). As compared to residents and controls, translocated animals demonstrated lower burrow switching during the year of translocation and also in the subsequent years (Fig. 2.5, Supplementary Fig. S5a). We did not find any differences between burrow switching levels of individuals exhibiting clinical signs of URTD and clinically healthy individuals
(χ²₁ = 2.51, P = 0.11). Burrow switching levels of all surveyed animals during drought years (indicated by lower winter rainfall), however, tended to be slightly lower in comparison to non-drought years (burrow switching: χ²₁ = 3.5, P = 0.06).

2.4 Discussion

Although direct social interactions among solitary species are relatively infrequent, individual preference for certain shared refuge and foraging spaces may lead to a highly structured social system [93]. In such species, knowledge of social network structure formed through refuge or forage associations can identify key influential individuals [30, 93], and provide early-warning signals for environmental (or anthropogenic) disturbances [57, 58] that may ultimately affect population fitness. In this study, we infer social associations between individuals of a relatively solitary species, the desert tortoise, by their asynchronous use of burrows. While descriptive approaches are common in the field of animal social networks [94], we sought to gain a mechanistic understanding behind individual variation in burrow-use associations of desert tortoises. The degree of an individual in a bipartite network has biological and ecological importance as it indicates a decision to switch refuges. Refuge switching is associated with a tradeoff between the costs of increasing exposure to heat, predators, increased risk of infection, and the benefits of finding food and mates. The outcome of observed refuge switching patterns is important as theoretical models predict reduced survival of populations due to suboptimal refuge use decisions [95]. Modeling optimal burrow switching that maximizes fitness in desert tortoises is challenging as it is difficult to quantify fitness costs in a long-lived species. Our study instead provides an approach to build baseline models of burrow use patterns. Any large deviation to baseline levels may indicate lower survival, foraging, and reproductive success for tortoises and thus
burrow switching can serve as an immediate indicator of population stressors affecting long-term fitness consequences.

We show that social networks in desert tortoises formed due to burrow use preferences cannot be explained by random associations. In several wildlife systems, spatial constraints can play a large role in shaping social networks [96], and non-random associations may not be definitive evidence of social organization in a population. Desert tortoise social associations, however, were only moderately correlated to spatial distances, which corroborates earlier studies that report social organization in desert tortoises [62, 63]. In general, the social networks were also clustered (0.23-0.59) and modular (0.34 - 0.68). However, higher clustering coefficient values have been reported in other social species [e.g., 0.54-0.57 in bottlenose dolphins [97], 0.57-0.87 in guppies [98], 0.81 in squirrels [99], 0.57-0.67 in primates [100]] and even in a few relatively solitary species that have been studied [e.g., 0.7 in raccoons [101], 0.59 in brushtail possum [102]]. The low (but significant) clustering coefficient value in desert tortoise social networks suggests that they do not form tight social bonds as compared to other social wildlife species. In social species, the network structure is known to affect population stability [103] and resistance to disease outbreaks [10, 13, 40]. Modular social networks of desert tortoises in particular can have important implications in the spread and persistence of infections. For example, few connections between communities in a social network can effectively localize new infections to a few individuals. For chronic infections such as URTD, these pockets of infection, however, can serve as sources of re-infection to other uninfected communities, eventually leading to persistent infection across the entire population.

Our analysis of burrow use heterogeneity in desert tortoises reveals that the period of the year and density of burrows around an individual are the main drivers behind the individual’s burrow switching decision. Low burrow switching levels in tor-
toises during winter and summer months reflects reduced movement to avoid severe weather conditions [104]. Individuals visit more burrows in the months of May-June and September-October which coincides with high activity of nesting and mating in adults. Seasonal rainfall also influences burrow switching in desert tortoises. Tortoises use fewer burrows in high rainfall conditions in March-April months, which possibly reflects reduced activity during cold weather associated with spring storms. Infrequent summer rains, on the other hand, increase tortoise activity as individuals emerge from burrows to rehydrate [105, 106]. Our results of high burrow switching during summer rains (July-August) are consistent with these reports of increased activity. We also find that non-reproductive tortoises use fewer burrows than adults, which may reflect differences in costs and benefits associated with switching burrows. Leaving a refuge can present a greater risk to non-reproductives that are more vulnerable to predation [107], are prone to thermal stress due to their smaller size [108], and do not benefit from the mating opportunities gained by burrow switching. Indeed, previous studies have found juveniles forage closer to their burrows and minimize time spent out of burrows [108, 109, 110]. Future studies and management plans of desert tortoises may consider differences in burrow switching between different non-reproductive tortoises in order to mitigate increased predation risk by pervasive predators such as ravens.

While it has been shown that a small fraction of burrows in desert tortoises are visited by multiple animals [61, 64], the mechanisms behind burrow popularity were previously unknown. Our results suggest that popular burrows can be identified using certain burrow characteristics such as surrounding topographical variables and age. As true burrow age is often hard to determine, we demonstrate the use of historical survey data to estimate proxy age of burrows. Once identified, these popular burrows can be surveyed throughout the year as there is only a minor effect of time of the year and seasonal rainfall on burrow popularity. Knowledge of active and popular
refuges can have two important implications for the conservation and management of wildlife species. First, population density estimates usually rely on observations of animals located outside refuge space [111]. For species that spend most of the time in a year in a refuge, survey of popular refuges can augment the current survey methods to get a more accurate estimate of population density. Secondly, declines of popular refuges can indicate reduced social interactions and mating opportunities for individuals. Reduced refuge popularity can also be indicative of higher mortality risk [112] found higher mortality of desert tortoise in flat open areas where burrows, as our results indicate, are less popular compared to rough higher elevation sites. Active popular burrows can therefore be used (a) as sentinels of population health and (b) to identify critical core habitat for conservation and adaptive management of a wildlife species.

Of three potential population stressors that we included in our model (disease, drought, translocation), translocation caused a change in burrow switching behavior of desert tortoises. Although translocated animals are known to have high dispersal tendencies [76, 113] and hence are expected to encounter and use more burrows, we found translocated individuals use fewer unique burrows than residents. Our results are supported by evidence of translocated tortoises spending more time on the surface and taking shelter under vegetation rather than using burrows [114]. The use of fewer burrows coupled with high dispersal rates can increase exposure of translocated animals to thermal stress and dehydration, potentially increasing mortality. Therefore, to improve translocation success, a fruitful area of investigation for future research will be to determine potential causes of this change in burrow use behavior in translocated tortoises. We used winter rain as a proxy of drought conditions as the Western Mojave receives most of its annual rainfall during the months of November-February and is important for the availability of food for desert tortoises in the spring [66, 72]. Our
results show a slight (but not significant) reduction in burrow use by tortoises during drought years. Reduced burrow switching may correspond to smaller home-ranges of desert tortoises observed during drought years [66]. Low winter rainfall condition is also known to increase predation of desert tortoises due to diminished prey resources [112, 115]. Lower burrow use during drought years can be therefore a behavioral response of desert tortoises to avoid predation or to reduce energy expenditure and water loss in years of low resource availability [105]. Contrary to previous studies [77], we did not find any effect of disease on burrow use behavior, possibly because we could not distinguish severe clinical signs with milder forms in our data. Although there was no evidence of disease influencing burrow use behavior in the present study, we note that it is likely for burrow use behavior (and in particular the burrows themselves) to drive infectious disease patterns in desert tortoises either directly, through cohabitation instances, or indirectly, by serving as focal sites of social interactions.

2.5 CONCLUSIONS

Our study demonstrates non-random associations in desert tortoises based on refuge use patterns. We formulate statistical models of burrow switching and popularity of burrows to investigate the mechanisms including environmental, topographical, density factors, and population stressors behind refuge use preferences of desert tortoises. In combination, these models help infer the mechanisms behind heterogeneity in refuge use from the perspective of individuals as well as from the perspective of the refuges. This approach is particularly useful for species that are not overtly gregarious. For these species, refuge switching often correlates to reproductive and foraging success, and patterns of refuge use can be an important aspect to consider before implementing any management or conservation strategy. For example, popular
refuges can be used to identify core habitat areas. In addition, sudden changes in the refuge switching behavior of individuals can be used as an early warning signal of disturbances that may ultimately affect population fitness. More broadly, our study provides insights towards the presence of and mechanisms behind non-random social structure and individual variation in a relatively solitary species by analyzing refuge-based associations. The structure of networks in social species is known to affect population stability and resilience to infectious diseases. Future studies are needed to establish such functional roles of social networks in relatively solitary species.
3.1 Introduction

Network analysis and modeling is a rapidly growing area which is moving forward our understanding of biological processes. Networks are mathematical representations of the interactions among the components of a system. Nodes in a biological network usually represent biological units of interest such as genes, proteins, individuals, or species. Edges indicate interaction between nodes such as regulatory interaction, gene flow, social interactions, or infectious contacts [116]. A basic model for biological networks assumes random mixing between nodes of the network. The network patterns in real biological populations, however, are typically more heterogeneous than assumed by these simple models [14]. For instance, biological networks often exhibit properties such as degree heterogeneity, assortative mixing, non-trivial clustering coefficients, and community structure (see review by Proulx et al.[116]). Of particular interest is community structure, which reflects the presence of large groups of nodes that are typically highly connected internally but only loosely connected to other groups [80, 117]. This pattern of large and relatively dense subgraphs is called assortative community structure. In empirical networks, these groups, also called modules or

---

communities, often correspond well with experimentally-known functional clusters within the overall system. Thus, community detection, by examining the patterns of interactions among the parts of a biological system, can help identify functional groups automatically, without prior knowledge of the system’s processes.

Although community structure is believed to be a central organizational pattern in biological networks such as metabolic [17], protein [18, 21], genetic [19], food-web [20, 23] and pollination networks [22], a detailed understanding of its relationship with other network topological properties is still limited. In fact, the task of clearly identifying the true community structure within an empirical network is complicated by a multiplicity of community detection algorithms, multiple and conflicting definitions of communities, inconsistent outcomes from different approaches, and a relatively small number of networks for which ground truth is known. Although node attributes in empirical networks (e.g., habitat type in foodwebs) are sometimes used to evaluate the accuracy of community detection methods [118], these results are generally of ambiguous value as the failure to recover communities that correlates with some node attribute may simply indicate that the true features driving the network’s structure are unobserved, not that the identified communities are incorrect.

A more straightforward method of exploring the structural and functional role of a network property is to generate graphs which are random with respect to other properties except the one of interest. For example, network properties such as degree distribution, assortativity and clustering coefficient have been studied using the configuration model [119], and models for generating random graphs with tunable structural features [14, 15, 120]. These graphs serve to identify the network measures that assume their empirical values in a particular network due to the particular network property of interest. In this work, we propose a model for generating simple,
connected random networks that have a specified degree distribution and level of community structure.

Random graphs with tunable strength of community structure can have several purposes such as: (1) serving as benchmarks to test the performance of community detection algorithms; (2) serving as null models for empirical networks to investigate the combined effect of the observed degrees and the latent community structure on the network properties; (3) serving as proxy networks for modeling network dynamics in the absence of empirical network data; and (4) allowing for the systematic study of the impact of community structure on the dynamics that may flow on a network. Among these, the use of random graphs with tunable strength of community structure to serve as benchmarks has received the most attention and several such models have been proposed [121, 122, 123, 124, 125, 126]. A few studies have also looked at the role of community structure in the flow of disease through contact networks [127, 128, 129, 130]. However, the use of modular random graphs, which can be defined as random graphs that have a higher strength of community structure than what is expected at random, is still relatively unexplored in other applications.

3.1.1 Previous work

In 2002, Girvan and Newman proposed a simple toy model for generating random networks with a specific configuration and strength of community structure [80]. This model assumes a fixed number of modules each of equal size and where each node in each module has the same degree. In this way, each module is an Erdős-Rényi random graph. To produce modular structure, different but fixed probabilities are used to produce edges within or between modules. Although this toy model has been widely used to evaluate the accuracy of community detection algorithms, it has limited relevance to real-world networks, which are generally both larger and much more
heterogeneous. Lancichinetti et al [121] introduced a generalization of the Girvan-Newman model that better incorporates some of these features, e.g., by including heterogeneity in both degree and community size. However, this model assumes that degrees are always distributed in a particular way (like a power law), which is also unrealistic. (A similar model by Bagrow [122] generates modular networks with power law degree distribution and constant community size.)

Yan et al [128] used a preferential attachment model to grow scale-free networks comprised of communities of nodes whose degrees follow a power-law distribution. And, models for special graph types such as hierarchical networks [123], bipartite networks [126], and networks with overlapping modules [125] have also been proposed. These models also make strong assumptions about the degree or community size distributions, which may not be realistic for comparison with real biological networks. A recently proposed model [124] does generate networks with a broad range of degree distributions, modularity and community sizes, but its parameters have an unclear relationship with desired properties (such as degree distribution and modularity), making it difficult to use in practice. Thus, while these models may be sufficient for comparative evaluation of community detection algorithms, they are of limited value for understanding their performance and output when applied to real-world networks.

An alternative approach comes from probabilistic models, of which there are two popular classes. Exponential random graph models (ERGMs) have a long history of use in social network analysis, and can generate an ensemble of networks that contain certain frequencies of local graph features, including heterogeneous degrees, triangles, and 4-cycles [131]. However, many classes of ERGMs exhibit pathological behavior when parameterized with triangles or higher-order structures [132], which severely limits their utility. Stochastic block models (SBMs) are more promising, but require a large number of parameters to be chosen before a graph can be generated. In this
approach, the probability of each link depends only on the community labels of its endpoints. Thus, to generate a network, we must specify the number of communities $K$, their sizes (in the form of a labeling of the vertices), and the $\binom{K}{2}$ (in the undirected case) group-pair probabilities. The result is a random graph with specified community sizes, where each community is an Erdős-Rényi random graph with a specified internal density, and each pair of communities is a random bipartite graph with specified density. The degree distributions of these networks is a mixture of Poisson distributions, which can be unrealistic. A recent generalization of the SBM due to Karrer and Newman [133] allows the specification of the degree sequence, which circumvents this limitation but introduces another set of parameters to be chosen. Although the stochastic block models can in principle be used to generate synthetic networks, they are more commonly used within an inferential framework in which community structure is recovered by estimating the various parameters directly from a network. As a result, the practical use of the SBM as a null model, either for general benchmarking of community detection algorithms or for understanding the structure of biological networks, remains largely unexplored, and we lack clear answers as to how best to sample appropriately from its large parameter space in these contexts. The SBM also does not provide a simple measure of the level of modularity in a network’s large-scale structure, which makes its structure more difficult to interpret. The SBM is a promising model for many tasks, and adapting it to the questions we study here remains an interesting avenue for future work.

3.1.2 Our approach

Here, we develop and implement a simple simulation model for generating modular random graphs using only a small number of intuitive and interpretable parameters. Our model can generate graphs over a broad range of distributions of network degree
and community size. The generated graphs can range from very small ($< 10^2$) to large ($> 10^5$) network sizes and can be composed of a variable number of communities. In Methods below, we introduce our algorithm for generating modular random graphs. In Results, we consider the performance of our algorithm and structural features of our generated graphs to show that properties such as degree assortativity, clustering, and path length remain unchanged for increasing modularity. We next demonstrate the applicability of the generated modular graphs to test the accuracy of extant community detection algorithms. The accuracy of community detection algorithms depends on several network properties such as the network mean degree and strength of community structure, which is evident in our analysis. Finally, using a few empirical biological networks, we demonstrate that our model can be used to generate corresponding null modular graphs under two different models of randomization. We conclude the paper with some thoughts about other applications and present some future directions.

### 3.2 Materials and methods

We present a model that generates undirected, simple, connected graphs with prescribed degree sequences and a specified level of community structure, while maintaining a graph structure that is otherwise as random (uncorrelated) as possible. Below, we introduce some notation and a metric for measuring community structure, followed by a description of our model and the steps of the algorithm used to generate graphs with this specified structure.
3.2.1 Measure of community structure

We begin with a graph $G = (V, E)$ that is comprised of a set of vertices or nodes $V(G) = \{v_1, \ldots, v_n\}$ and a set of edges $E(G) = \{e_1, \ldots, e_m\}$. $G$ is undirected and simple (i.e. a maximum of one edge is allowed between a pair of distinct nodes, and no “self” edges are allowed). The number of nodes and edges in $G$ is $|V(G)| = n$ and $|E(G)| = m$, respectively. The neighborhood of a node $v_i$ is the set of nodes $v_i$ is connected to, $N(v_i) = \{v_j \mid (v_i, v_j) \in E, v_i \neq v_j, 1 \leq j \leq n\}$. The degree of a node $v_i$, or the size of the neighborhood connected to $v_i$, is denoted as $d(v_i) = |N(v_i)|$. A degree sequence, $D$, specifies the set of all node degrees as tuples, such that $D = \{(v_i, d(v_i))\}$ and follows a probability distribution called the degree distribution with mean $\bar{d}$.

Each community or module $C_k$ is defined as a subset of $G$ that contains both nodes, $V(C_k)$ and edges $E(C_k)$, where both the endpoints of each edge in $E(C_k)$ are contained in $V(C_k)$. $K$ is the number of modules in $G$ and $k \in [1, K]$. Each node $v_i$ of $G$ has a within-degree, $d_w(v_i) = |N(v_i) \cap V(C_k)|$, which is the number of within-edges connecting $v_i$ to other nodes of the same module $C_k$; and a between-degree, $d_b(v_i) = |N(v_i) - V(C_k)|$, i.e. the number of between-edges connecting $v_i$ to nodes in different modules (here, the minus operator represents set difference). The strength of the community structure defined by a partition, $\{C_k\}$, can be measured as modularity, $Q$, and is defined as

$$Q = \sum_{k=1}^{K} (e_{kk} - a_k^2)$$

(3.1)

where $e_{kk} = \frac{|E(C_k)|}{|E(G)|}$ denotes the proportion of all edges that are within module $C_k$, and

$$a_k = \frac{\sum_{v_i \in C_k} d(v_i)}{2|E(G)|}$$

represents the fraction of all edges that touch nodes in community $C_k$. When $Q = 0$, the density of within-community edges is equivalent to what is expected when edges are distributed at random, conditioned on the given
degree sequence. Values approaching \(Q = 1\), which is the maximum possible value of \(Q\), indicate networks with strong community structure. Typically, values for empirical network modularity fall in the range from about 0.3 to 0.7 [134]. However, in theory Good et al. [37] show that maximum \(Q\) values depend on the network size and number of modules.

In order to generate a graph with a specified strength of community structure, \(Q\), equation (1) represents our first constraint, which we rewrite below in terms of the expected value of \(Q\), (full derivation in Additional file 1):

\[
E[Q] = \sum_{k=1}^{K} \left[ \frac{d_{w,s_k} m}{m} - \left( \frac{d_{s_k} m}{m} \right)^2 \right]
\]

(3.2)

where \(d_w\) and \(d\) are the average within-degree and average degree, respectively, and \(s_k = |V(C_k)|\) is the module size for module \(k\). Thus, equation (2) allows us to specify \(d_w\) in terms of \(Q\), \(d\), \(m\) and \(s_k\), assuming that the module-specific average degree and average within-degree are equal to \(d\) and \(d_w\), respectively. When \(s_k = \bar{s}\) for all \(k\), \(E[Q]\) reduces to \(d_w - 1/K\).

We note that as the average within-degree \((d_w)\) approaches the average degree \((d)\), the graph, \(G\) becomes increasingly modular. Hence, the maximum modularity for \(G\) with \(K\) modules can be estimated as:

\[
Q_{\text{max}} \simeq \sup(Q) = 1 - \frac{1}{K}
\]

(3.3)

3.2.2 ALGORITHM

We present a model and an algorithm that generates undirected, unweighted, simple and connected modular random graphs. The model is specified by a network size \((n)\), degree distribution \((p_d)\), an expected modularity \((E[Q])\), the number of modules \((K)\), and the module size distribution \((P(s))\), with mean \(\bar{s}\). (We note a degree sequence,
$d(v_i)$, may be specified instead of a degree distribution, $p_d$). The algorithm proceeds in four steps:

1. Assign the $n$ network nodes to $K$ modules based on the size distribution $P(s)$.

2. Assign degrees, $d(v_i)$, to each node $v_i$ based on $p_d$ and $\bar{d}$. We next assign within-degrees, $d_w(v_i)$, to each node $v_i$ by assuming that the within-degrees follow the same distribution as $p_d$ with mean $\bar{d_w}$, which is estimated based on equation (2) above (Figure 3.1a).

3. Connect between-edges based on a modified Havel-Hakimi model and randomize them. (Figure 3.1b)

4. Connect within-edges based on the Havel-Hakimi model and randomize them. (Figure 3.1c and 1d)

The generated graph then has a degree distribution that follows $p_d$ with mean $\bar{d}$, $K$ modules with sizes distributed as $P(s)$, and a modularity $Q \approx E[Q]$. We set an arbitrary tolerance of $\epsilon = 0.01$, such that the achieved modularity is $Q = E[Q] \pm \epsilon$. The graph is also as random as possible given the constraints of the degree and community structure, contains no self loops (edges connecting a node to itself), multi-edges (multiple edges between a pair of nodes), isolate nodes (nodes with no edges), or disconnected components. Below, we elaborate on each of the steps of this algorithm.

**Assigning Nodes to Modules**

We sample module sizes, $s_k$, for each of the $K$ modules from the specified module size distribution, $P(s)$ so that $\sum s_k = n$. The $n$ nodes are then arbitrarily (without loss of generality) assigned to each module to satisfy the sampled module size sequence.
Figure 3.1: Schematic representation of the steps of random modular network generator. (a) The algorithm assigns a within-degree and between-degree to each node, which are represented here as half-within-edges and half-between-edges respectively. (b) The half-between-edges are then connected using a modified version of the Havel-Hakimi algorithm, and to remove degree correlations, the between-edges are randomized. (c) Finally, the half-within-edges are connected using the standard Havel-Hakimi algorithm for each module and (d) the within-edges are randomized to remove degree correlations.
ASSIGNING DEGREES

Based on the degree distribution specified, a degree sequence is sampled from the distribution to generate a degree, \(d(v_i)\), for each node \(v_i\) (unless a degree sequence is already specified in the input). To ensure that the degree sequence attains the expected mean of the distribution (within a specified threshold) and is realizable, we verify the Handshake Theorem (the requirement that the sum of the degrees be even) and the Erdős-Gallai criterion (which requires that for each subset of the \(k\) highest degree nodes, the degrees of these nodes can be “absorbed” within the subset and the remaining degrees) [135], and that no node is assigned a degree of zero.

Unless a within-degree sequence is specified, we assume that the within-degree distribution follows the class of the degree distribution specified, \(p_d\), with mean \(\overline{d_w}\) based on equation (2) (i.e. a generated network with a Poisson degree distribution of mean \(\overline{d}\) also has a Poisson within-degree distribution with mean \(\overline{d_w}\)). This assumption is considered reasonable as it holds true for several of the empirical networks we analyze (shown in Figure S1 in Additional file 1). However, our model can be extended for arbitrary within-degree distributions (or sequences) (see Table S1 in Additional file 1), although the space of feasible within-degree distributions given a degree distribution is restricted. Next, we sample a within-degree sequence, \(d_w(v_i)\), from this within-degree distribution. Using rejection sampling, we ensure that the within-degree sequence attains the expected overall mean, \(\overline{d_w}\) within a tolerance \(\epsilon_{d_w} = \epsilon \overline{d}\) (with details in the Additional file 1), and satisfies the following conditions:

- **Condition 1**: \(d(v_i) \geq d_w(v_i)\) for all \(v_i\). To ensure this, we sort the degree sequence and within-degree sequence, independently. If \(d(v_i) < d_w(v_i)\) for any \(v_i\) in the ordered lists, the condition is not satisfied. In Figure S2 of Additional file 1,
we discuss the rejection rates for the rejection sampling of both the degree and within-degree sequence.

- Condition 2: a realizable within-degree sequence for each module, \( C_k \), as defined by the Handshake Theorem and the Erdos-Gallai criterion.

In addition, to ensure that each module approximately achieves the overall mean within-degree, \( \bar{d}_w \), we specify the following constraint: \( \max\{d_w(v_i)\}_{v_i \in G} \leq \min[s_k] \). If the sampled module sizes do not satisfy this criteria, the module sizes are re-sampled or an error is generated.

The between-degree sequence is generated by specifying \( d_b(v_i) = d(v_i) - d_w(v_i) \) for each node \( v_i \). To test if the between-degree sequence is realizable, we impose a criterion developed by Chungphaisan [136] (reviewed by Ivanyi [137]) for realizable degree sequences in multigraphs. To do so, we imagine a coarse graph, \( H \), where the modules of \( G \) are the nodes of \( H \) (i.e. \( V(H) = \{C_1, C_2, \ldots, C_K\} \)), and the between-edges that connect modules of \( G \) are the edges of \( H \). We note that \( H \) is a multigraph, because \( G \) allows multiple between-edges of \( G \) to connect each pair of modules. In this case, the degree sequence of \( H \) is

\[
D = \{(C_k, d(C_k)) | d(C_k) = \sum_{v_j \in C_k} d_b(v_j), k = 1 \ldots K}\.
\]

The Chungphaisan criterion then specifies that the multigraph degree sequence \( \{d(C_k)\} \) on \( H \) is realizable if the following conditions are satisfied:

- Condition 1: the Handshake theorem is satisfied for \( \{d(C_k)\} \): \( \sum_{k=1}^{K} d(C_k) = \sum_{k=1}^{K} \sum_{v_j \in C_k} d_b(v_j) \) is even

- Condition 2: \( \sum_{k=1}^{j} d(C_k) - bj(j-1) \leq \sum_{k=j+1}^{K} \min[jb, d(C_k)] \) for \( j = 1, \ldots, K - 1 \).
Here, $b$ is defined as the maximum number of edges allowed between a pair of nodes in $H$; in our case, $b = \max\{d_b(v_i)\}$, the maximum between-degree of any node $v_i \in G$.

We also generate graphs with $Q = 0$ by assuming the network is composed of a single module with no between-edges. Thus, $d_w(v_i) = d(v_i)$ and $d_b(v_i) = 0$ for all $v_i \in G$.

**Connecting edges**

Based on the within-degree sequence and between-degree sequence specified above, edges are connected in two steps (Figure 3.1). Nodes that belong to different modules are connected based on their between-degree to form between-edges (Figure 3.1b) and nodes that belong to the same module are connected according to their within-degree to form within-edges (Figure 3.1c and 3.1d).

We connect between-edges using a modified version of the Havel-Hakimi algorithm. The Havel-Hakimi algorithm [138, 139] constructs graphs by sorting nodes according to their degree and successively connecting nodes of highest degree with each other. After each step of connecting the highest degree node, the degree list is resorted and the process continues until all the edges on the graph are connected. Here, we modify this to construct between-edges by sorting nodes by highest between-degree, in order of highest total between-degree for the module to which they belong, and successively connecting the node at the top of the list randomly with other nodes. Connections are only made between nodes if they are not previously connected, belong to different modules, and do not both have within-degree of zero (to avoid disconnected components). After each step the between-degree list is resorted, and the process continues until all between-edges are connected. After all between-edges have been connected, the connections are randomized using a well-known method of rewiring.
through double-edge swaps [140]. Specifically, two randomly chosen between-edges \((u, v)\) and \((x, y)\) are removed, and replaced by two new edges \((u, x)\) and \((v, y)\), as long as \(u\) and \(x\), and \(v\) and \(y\) belong to different modules, respectively. The swaps are constrained to avoid the formation of self loops and multi-edges. This process is repeated a large number of times to randomize edges.

We then connect within-edges using the standard Havel-Hakimi algorithm, applied to each module independently. Specifically, within-edges of a module are connected by sorting nodes of the module according to their within-degree and successively connecting nodes of highest within-degree with each other. Connections are only made between nodes if they are not previously connected, and do not both have a between-degree of zero (to avoid disconnected components). After each step the within-degree list is resorted and the process continues until all the within-edges of the module are connected. The connections are then randomized by rewiring through double-edge swaps [140]. We do not specify that each module be connected (only that the full graph is connected). However, if this is required, Taylor’s algorithm can be used to rewire pairs of edges until the module is connected [141]. Specifically, the algorithm selects two random edges \((u, v)\) and \((x, y)\) that belong to two different disconnected components of the module. As long as \((u, x)\) and \((v, y)\) are not existing edges, the \((u, v)\) and \((x, y)\) edges are removed and \((u, x)\) and \((v, y)\) are added. Taylor’s theorem proves that following such operation any disconnected module can be converted to a connected module with the same degree sequence.

3.3 Results and discussion

Using our simulation algorithm, we were able to generate modular random graphs of variable network size, number of communities, degree distribution, and community
size distribution. In Figure 3.2, we show sample networks of varying levels of mod-
ularity, \( Q = 0.1, 0.3, 0.6 \). We note that a network with three modules can approach
a maximum modularity value of \( 2/3 \) (from equation 3), and thus \( Q=0.6 \) is a rela-
tively high modularity for this particular network type. In the sections that follow,
we consider the algorithm performance, as well as structural properties of the gener-
ated graphs. We then highlight two applications of our model: 1) to generate bench-
mark graphs for validation of community detection algorithms and 2) to generate
null graphs for the analysis of empirical networks. Community detection algorithms
assist in identifying community structure in empirical networks. Our model is able
to generate modular networks \textit{de novo} to test these algorithms. Once community
structure has been identified in an empirical network with a community detection
algorithm, the number of communities and the modularity level \( (Q) \) (and, if desired,
the community size distribution and within-degree sequence) can be used as input
to our model to generate graphs that can act as random controls to test hypotheses
about the empirical system.

3.3.1 Performance and properties of generated graphs

Performance

Our model generates graphs that closely match the expected modularity and degree
distribution. The deviation of the observed modularity is less than 0.01 from the
expected value, given the specified partition. The modular random graphs with
Poisson degree distribution generated by our model are similar to the ones described
by Girvan and Newman\cite{80} with linking \( (p_{in}) \) and cross-linking probability \( (p_{out}) \)
equal to \( \frac{\bar{d}_w}{\bar{d}-1} \) and \( \frac{\bar{d}_w-\bar{d}}{\bar{d}(K-1)} \) respectively. However, our model overcomes several limita-
tions of the model proposed by Girvan and Newman\cite{80} and others \cite{121, 122} by
Figure 3.2: Modular random graphs with $n=150$, $m=375$, $K=3$, $P(s=50)=1$ and $p_k$ is power law with modularity values of: a) $Q=0.1$; b) $Q=0.3$; and c) $Q=0.6$. As the modularity increases, the ratio of the total number of edges within modules to the number of edges in the network increases (i.e. $\bar{d}_w$ increases), while the remaining parameter values (degree distribution, network mean degree, number of modules) are held constant.

considering heterogeneity in total degree, within-module degree distribution, and module sizes. Unlike many of the existing models [123, 124, 125, 126], our model can generate modular random graphs with arbitrary degree distributions, including those obtained from empirical networks. Though we discuss modular random graphs with positive $Q$ values, our model can also generate disassortative modular random graphs (see Figure S3 in Additional file 1). In this case, nodes tend to connect to nodes in other modules and thus the density of edge connections within a module is less than what is expected at random. Additionally, we also compare our model to graphs generated based on a degree-corrected stochastic block model (SBM). The details of the parameterization of the SBM and the results are shown in Figure S4 in Additional file 1).
**Structural properties**

There are several other topological properties (besides degree distribution and community structure) that can influence network function and dynamics. The most significant of these properties are degree assortativity (the correlation between a node’s degree and its neighbor’s degrees), clustering coefficient (the propensity of a node’s neighborhood to also have edges among them) and average path length (the typical number of edges between pairs of nodes in the graph). We have developed this model to generate graphs with specified degree distribution and modularity, while minimizing structural byproducts. Thus, it is important to confirm that we have reached this goal with the generative model above.

To evaluate the status of other structural properties due to the generative model, we specify graphs of \( n = 2000 \) following Poisson \((\lambda^k e^{-\lambda}/k!)\), geometric \((p(1-p)^{k-1})\), and power-law \((k^{-\alpha}/\zeta(\alpha))\) degree distributions with \( \overline{d} = 10 \). We chose these particular types of degree distributions as they have widely studied in the context of biological networks[142, 143, 144]. Each network has \( K = 10 \) modules and a module size distribution \( P(s = 200) = 1 \). We generate modular random graphs with these specifications and modularity values that range from \( Q = 0 \) to \( Q = 0.8 \), in steps of 0.1. For each level of modularity, we generated 50 such modular random graphs and calculated the degree assortativity \((r)\), clustering coefficient \((C)\), and average path length \((L)\) for each network, which is illustrated in Figure 3.3. In networks with random community structure \((Q = 0)\), that is random graphs with specified degree distributions (such as those that would be generated by the configuration model [119]), the value of \( r, C, \) and \( L \) are what are expected at random. In Figure 3.3, we show that for increasing values of modularity, degree assortativity, clustering coefficient, and average path length remain relatively constant for all three network types (i.e. Poisson, geometric
Figure 3.3: Values of (a) assortativity, $r$, (b) clustering coefficient, $C$, and (c) path length, $L$ in modular random graphs do not vary significantly with increasing modularity ($Q$). Each graph has $n = 2000$ nodes, a mean degree $\bar{d} = 10$ and $K = 10$ modules with $P(s = 200) = 1$. The data points represent the average value of 50 random graphs. Standard deviations are plotted as error bars.

and power-law). At the highest levels of modularity, edge connections are constrained, particularly for the heavy-tailed geometric and power-law degree distributions, leading to an increase in clustering coefficient. Correlations between high clustering coefficient and high modularity have also been observed before [14]. The average path length for all network types also increases at the highest levels of modularity, likely reflecting the lack of many paths between modules, requiring additional steps to reach nodes in different modules. Thus, our model is able to increase levels of modularity in random graphs without altering other topological properties significantly.

Biological networks show remarkable variation in network size, connectivity and community size distribution, with some of them having particularly small network size, high degree, and small module sizes (e.g. food-web networks). We therefore tested the performance of our generated networks under deviations in the network specifications of size, mean degree and module size distribution (results presented in Additional file 1: Figure S5, S6 and S7). We find that the structural properties of our generated modular random graphs remain constant, except for two constraining
conditions: a) high average degree ($\bar{d} > 10$) and b) low average module size ($\bar{s} < 50$). At these parameter extremes, the modular random graphs become degree disassortative and have increased clustering coefficient. A similar observation of network degree disassortativity has been made in hierarchically modular networks [145]. In these two scenarios, the highest value of within-degree ($d_w(v_i)$) that a node can attain is constrained by the community size, which reduces the number of possible high within-degree nodes. As a consequence high within-degree nodes must connect to low within-degree nodes more than expected, resulting in a degree disassortative network. In these two cases, modules also become more dense and thus create more triangles resulting in a gradual increase in clustering. Path length, on the other hand, is not affected by these conditions and shows a consistent dependence on network size and mean degree, which is well known [146, 147].

3.3.2 APPLICATION: BENCHMARK GRAPHS FOR COMMUNITY-DETECTION ALGORITHMS

Detecting communities in empirical networks has been an area of intensive research in the past decade [148] since Girvan and Newman’s seminal paper on community detection [80]. Extant techniques such as modularity maximization, hierarchical clustering, the clique-based method, the spin glass method etc. aim at achieving high levels of accuracy in detecting the correct partition (for a detailed review see [148]), but have their own set of strengths and weaknesses. Choosing the best algorithm can be a difficult task especially as algorithms often use distinct definitions of communities and perform well within that description. Thus, it is exceedingly important to test community-detection algorithms against a suitable benchmark. We propose our modular random graphs as benchmark graphs for the validation of existing and new algorithms of community detection.
To illustrate this use, we test the performance of six popular community detection algorithms: the Louvain method [149], fast modularity method [150], the spin-glass based method [151], the infoMAP method [152], label propagation [153] and the random-walk based method [154] using our modular random graphs as benchmarks. Specifically, we generate a modular random graph for each level of modularity and used these community detection algorithms to detect their community structure. We also test the performance of the algorithms on random graphs of specified degree distribution, with no modularity (i.e. $Q=0$). Figure 3.4 summarizes the performance of these algorithms, as measured by the estimated $Q$, for modular random graphs with three different degree distributions (Poisson, geometric and power-law). We also investigated the robustness of these algorithms on replicate modular random graphs at each modularity level, with the results presented in (Additional file 1: Figures S8, S9 and S10).

The Louvain, fast modularity algorithm, random-walk and infoMAP algorithm overestimate the modularity for networks with weak community structure, and underestimate the modularity for networks of moderate community structure across all three network types (Figure 3.4). Spinglass and label-propagation consistently underestimate the modularity of both weak and moderate community structure. All the algorithms are fairly accurate at the highest strengths of community structure across the various network types. The accuracy at a particular level of modularity and degree-distribution, however, varies for different algorithms. For instance, the performance of spin-glass algorithm is better for Poisson modular random graphs at modularity values of 0.5-0.6, whereas the Louvain and label-propagation algorithm out-perform on geometric random modular graphs at these modularity values.

In addition to comparing the estimated values of modularity to the known values in the modular random graphs, we can compare the similarity in the partitions detected
Figure 3.4: Performance of the Louvain method [149], fast modularity method [150], the spin-glass based method [151], the infoMAP method [152], label propagation [153] and the random-walk based method [154] in networks with mean degree 10. Fill circles, open circles and triangles represent networks with Poisson, geometric and power-law degree distributions, respectively. Each data point represents the average over ten modular random graphs. Error bars represent standard deviations. The solid line is the reference line where estimated modularity is equal to the input modularity.
by the algorithms to the true partitions. For this comparison, we use the Jaccard similarity \((J)\), which measures the similarity between two partitions based on the proportion of the union of the partitions that is made up by the intersection of the partitions \([155]\); as well as the Variation of Information \((VI)\), which measures the distance between two partitions based on the amount of information lost when going from one partition to another \([156]\). These results are presented in the (Additional file 1: Figure S11). As reflected in the results above, we find that partitioning is inaccurate when the true community structure is weak but improves as the \(Q_{true}\) value increases. These observations have also been noted before by Lacichinetti and Fortunato \([157]\).

3.3.3 Application: Null analysis of empirical networks

It is crucial to have random controls in the study of biological systems. Our algorithm can be used to generate null models and applied to the detection of structure in empirical biological networks. These null networks can be used to test hypotheses regarding the role of modularity and other topological features of the empirical networks. To do so, one would first determine the number of communities and modularity level \((Q)\) of the sampled network using an appropriate community detection algorithm (the previous section describes the use of random modular graphs to validate existing algorithms of community detection). Our algorithm can then be used to generate an ensemble of networks that match the empirical degree structure and community structure, and then compare the structural, functional, or dynamical properties of the empirical network to those of the generated modular random graphs. Because our model generates graphs without any structural byproducts (as illustrated in a previous section), this is an appropriate model for generation of null models. We note that our algorithm does not necessarily require knowledge of the complete empirical
network, but rather only estimates of the degree structure and community structure. The literature on algorithms for inference of network structure from a sample is growing, and currently includes work on inference of missing nodes, edges and even community structure [158, 159, 160].

We demonstrate this application using four classes of biological networks, namely: a) a food-web, representing the trophic interactions at Little Rock Lake in Wisconsin with a network size of 183 and average degree = 26.8 [161]; b) a protein-protein interaction network in *Saccharomyces cerevisiae* (a yeast) of size= 4713 and average degree = 6.3[162]; c) a metabolic interaction network of *Caenorhabditis elegans* of size = 453 and average degree 9.0 [163]; and d) a network of social interactions in a community of dolphins living off Doubtful Sound, New Zealand of size = 62 and average degree = 5.1 [164]. Visualizations of the dolphin social interaction network and the food-web trophic interaction network and its modular random counterpart are shown in Figure 3.5.

For each of these four empirical networks, we generate modular random graphs (Figure 3.6, light gray bars) with three parameters estimated from the empirical networks: (a) the degree sequence, $p_k$ (b) the modularity, $Q$ and (c) the average community size, $\bar{s}$. We note that as our goal is to construct null models, we assume that communities are of equal size, i.e. $P(\bar{s}) = 1$, and that the within-degree distribution matches the degree distribution fitted from the specified degree sequence (with estimated mean, $\overline{d_w}$). (A second class of null models can be constructed with $P(s)$ and the within-degree sequences estimated from the empirical networks, and we do this in Table S2 of Additional file 1). Specifically, we generate 25 such random graphs and measure structural properties of the generated graphs including clustering coefficient ($C$), average path length ($L$), degree assortativity ($r$).
Figure 3.5: Visualization of empirical and random graphs of social interaction of dolphins and food-web trophic interactions at the Little Rock Lake in Wisconsin. Figure (a) is the empirical network of Dolphin social network, (b) its modular random graph, and (c) its random graph counterpart with matched degree distribution \((Q=0)\). Figure (d) is the empirical network for the food-web trophic interaction at Little Rock Lake in Wisconsin, (e) is its modular random graph and (f) its random graph counterpart with matched degree distribution. Modular random graphs have generated to match the overall degree distribution, network mean degree, the level of modularity and the number of modules of the empirical graphs. Random graphs with matched degree distribution are based on the configuration model.
Figure 3.6: Comparisons of empirical networks, modular random graphs and random graphs with matched degree distribution (based on the configuration model). The figure summarizes network statistics of the empirical network as well as the ensemble mean of two types of random graphs in terms of (a) modularity, $Q$; (b) assortativity, $r$; (c) path length, $L$ and (d) clustering coefficient, $C$. The path length value for the empirical Yeast-Protein interaction network is missing as the network contains disconnected components. Error bars denote standard deviation from the ensemble mean of the generated random graphs. Errors bars for modular random graphs in Figure 3.6(a) have been omitted as the value of modularity ($Q$) match the empirical networks perfectly. FW= Little Rock food web, YP= Yeast protein interaction network, CM= C.elegans metabolic network and DS= Dolphin social network.
We also generate random graphs based on the configuration model that have the same degree distribution and average network degree as the empirical network but are random with respect to other network properties for each of the four empirical networks (Figure 3.6, dark gray bars). Our modular random graph model identifies which network measures assume their empirical values in a particular network because of (i) the observed degrees and (ii) the latent community structure. The configuration model, on the other hand, only specifies (i) and not (ii) [119]. Comparison to these configuration model networks thus helps us highlight the utility of our model to identify which empirical patterns in a network are deserving of further investigation. Figure 3.6 shows the value of each of these properties for the empirical networks as well as the ensemble mean of modular random and random graphs with matched degree distribution.

From Figure 3.6 it is evident that none of the empirical biological networks have network structure identical to their null counterparts. This suggests that the structure of each of these biological systems is governed by more than what is specified by the degree distribution and community structure. However, the observed network properties of empirical networks are closer to the ensemble means of the modular random graphs, which indicates that modularity is an essential structural component of real biological networks and that it plays an important role in influencing other structural properties of the network. For instance, compartmentalization induced by modularity promotes species persistence and system robustness by containing localized perturbation [22, 165, 166], which might favor their selection during the course of evolution. Our results show that the empirical networks tested have a much higher modularity than the simple random graphs (Figure 3.6a) and therefore provide evidence for this selection. Out of the three network properties that we tested apart from modularity, we found clustering coefficient of the generated random graphs to
be significantly different from each of the empirical counterparts. This may point to a functional role for “triangles” in these biological networks, significantly above or below what is prescribed by the degree and community structure.

Little Rock Lake food web interactions (FW): Among the four empirical networks that we tested, the properties of the ensemble mean of null models such as assortativity and path length closely match most of the observed properties of Little Rock food web. The observed clustering coefficient of food web is strikingly lower than either of the random graphs which confirm the observations of low clustering in food web made by earlier studies (Figure 3.6d). The observed path length of this food web is short (Figure 3.6c) and only slightly longer than the path lengths of random graphs, which has also been noted before [167, 168, 169]. We note that for this food web, the structural properties of the random graphs with matched degree distribution are quite similar to those of modular random graph counterparts, suggesting that the degree distribution, particularly the high density of edges in the network governs most of the other topological characteristics of this network. Modularity, on the other hand, seems to play a minor role in dictating the structural properties of this network.

Yeast protein-protein interaction network (YP): The empirical yeast protein network is more disassortative than the ensemble mean of null modular graphs (Figure 3.6b). Disassortative interactions in protein-protein interaction networks are known to reduce interferences between functional modules and thus increase the overall robustness of the network to deleterious perturbations [18], while also allowing for functions to be performed concurrently [170]. The results therefore suggest that disassortative interactions may be selected for in the evolution of biological networks. From Figure 3.6(d) it is also evident that the yeast protein network has a higher
value of clustering coefficient than the expected value predicted by the modular random graphs. A high value of clustering coefficient indicates that there are several alternate interaction paths between two proteins, making the system more robust to perturbation[171].

*C.elegans* metabolic interaction network (CM): The *C.elegans* metabolic network demonstrates a shorter path length but higher clustering coefficient than both modular and random graphs with matched degree distribution (Figure 3.6c and 3.6d). A high clustering coefficient and short path length suggests that the graph has small-world properties, which has been observed in other metabolic networks as well [172]. A highly disassortative degree structure is also well known in metabolic networks, although the mechanism leading to this property is unclear (see review by [142]). As the predicted value of disassortativity of the modular random graphs is closer to the observed value, our results suggest that the strong community structure of the metabolic networks could be one of the factors contributing to high degree disassortativity. (As discussed earlier, community structure leads to significant degree correlations in small networks with long-tailed degree distributions; see Figure S5 in Additional file 1 for an example).

Social interaction network of dolphins network at Doubtful Sound, New Zealand (DS): The empirical social interaction network of dolphins that we investigated demonstrated a negative assortativity (or disassortativity) similar to other real biological networks (Figure 3.6b). Interestingly, the assortativity value of both null modular and random graphs with matched degree distribution counterparts of the dolphin network is lower than the observed value, which suggests that the network is more assortative than expected. Degree assortativity has also been observed in other animal [173] and
human [117] social interaction networks. This result is quite intuitive for a social network and is also referred to as homophily: more gregarious individuals tend to interact with other gregarious individuals while introverted individuals prefer to associate with other introverts[15]. The empirical dolphin network also demonstrated a lower value of clustering coefficient than the expected values of either null model. Low clustering coupled with high degree assortativity indicates that dolphin populations may be more susceptible to the propagation of infection or information, as transmission may occur rapidly through the entire network with such properties [173, 174].

3.4 Conclusions

In summary, the model that we propose in this study generates modular random graphs over a broad range of degree distribution and modularity values, as well as module size distributions. We highlight that our model is specifically designed to generate networks which have modularity evenly divided across its modules, modulo the impact of module size. This means that we are mitigating the resolution limit effect and indeed generating networks with the maximum modularity partition. We also confirm that structural properties of our generated modular graphs such as assortativity, clustering and path length remain unperturbed for a broad range of parameter values. This important feature allows these graphs to act as benchmark and control graphs to explicitly test hypotheses regarding the function and evolution of modularity in biological systems. Of the approaches available, our method provides flexibility and has been explored the most fully for these applications.

Compartmentalization of biological networks has been an area of great interest to biologists. What we refer to as community structure in this work is any segregation of a biological system into smaller subunits inter-connected by only a few connections.
It has been suggested that modularity in a system promotes system robustness and enhances species persistence by containing localized perturbations [22, 166]. Metabolic networks of organisms living in a variable environment have indeed been found to be more modular [165]. Maintaining and selecting for modularity in biological networks, however, comes at a great cost of reducing system complexity [175], longer developmental time and cost of complete module replacement in case of failure [176]. It is therefore unclear why modularity would be strongly selected for as a structural feature of biological systems. There is also a lack of evidence to prove that the functional localization of sub-goals overlaps with the structural segregation of the network into community structure. Our work provides a tool for the systematic study of network structure (through benchmark graphs) and of the impact of connectivity and compartmentalization on system function and dynamics (through control graphs).

The detection of community structure plays a crucial role in our topological understanding of complex networks. Currently the performance of community detection methods is usually evaluated based on ground-truth from real networks. However, determining reference communities in real networks is often a difficult task. Also, ground truth data on empirical network partitions do not necessarily identify system features based on network topology and thus may create a bias when analyzing community structure. A more convenient technique of evaluating community detection method is to use artificial random graphs, but has been limited as most of the models fail to incorporate degree heterogeneity of real networks. By providing a systematic method to generate benchmark graphs, our model can aid in the development of more robust community detection algorithms, and therefore improve our topological understanding of empirical networks.

A step beyond identifying the topological presence of network communities is the understanding of its evolution as well as the functional and dynamical role of com-
community structure. We believe this process can be facilitated by using an appropriate class of control or null graphs. As a model for generating null networks, our method joins a suite of random graph models, each contributing to a hierarchy of null models. The simplest model for generating random graphs (based on only a single parameter) is the Erdős-Rényi random graph model, which produces graphs that are completely defined by their average degree and are random in all other respects. A slightly more complex and general model is one that generates graphs with a specified degree distribution (or degree sequence) but are random in all other respects [119, 177, 178]. These models can be extended to sequentially include additional independent structural constraints, such as degree distribution and clustering coefficient [14], or degree structure and community structure, as we have demonstrated here. A further extension to this work will be designing models that generate random graphs with multiple structural constraints. For example, our model can be combined with the one proposed by [14] to generate random graphs with specified degree distribution as well as tunable strength of modularity and clustering coefficient.

3.5 Availability and requirements

**Project name:** Modular random graph generator

**Project home page:** http://github.com/bansallab

**Operating system(s):** Platform independent

**Programming language:** Python 2.7

**Other requirements:** Networkx Python package

**License:** BSD-style

**Any restrictions to use by non-academics:** None
Chapter 4

Unraveling the disease consequences and mechanisms of modular structure in animal social networks

4.1 Introduction

Social behavior is ubiquitous in vertebrates and arthropods. An evolutionary product of sociality in animal populations is group living. Group living societies benefit from improved success in inter-group competition, protection from predation, mating opportunities, cooperative care of young, foraging success and group defense. On the other hand, group living is also associated with the costs of elevated disease burden due to higher frequency of contact between hosts [39]. Recently, it has been suggested that modular subdivisions in social interaction of species alleviates the association between group size and disease burden [35, 179].

Modular organization emerges in animal groups when subsets of conspecifics consistently interact with each other more often than they do with other individuals in the group, forming subgroups. Such modular structure is widespread across social networks in wildlife species, including relatively solitary species [180, 181], and can emerge from a combination of social behavior, demography, environmental and landscape factors. However, theoretical investigations of the impact of modularity on spreading processes have produced mixed results in the past. While some studies

---

suggest modular subdivisions reduce transmission by creating structural bottlenecks [26, 130, 182], a few have found modular structure to increase outbreak size due to higher connectivity within subgroups [183, 184]. This discrepancy in past research implies the need to consider realistic ranges of modularity (as not all animal social networks exhibit the high levels of modular subdivision typically investigated in past studies), and the need to mechanistically understand how modular structure in social networks affects the dynamics of disease spread.

Thus, we propose an empirically-driven investigation of the structural features of interaction patterns that are associated with modular subdivisions in animal groups, including the ones that change the degree of network fragmentation (i.e., the number of subgroups), promote subgroup cohesion (i.e., individual preferences to interact with subgroup members), and drive variation in subgroup size (Fig. 4.1).

We have three objectives: 1) identify features of animal interaction patterns that increase the strength of modular organization in their social networks, 2) explore the conditions under which modular organization influences the dynamics of disease spread and explain the mechanisms behind observed effects, and 3) investigate whether knowledge of modular organization in animal social networks is necessary (and sufficient) to build accurate contact network models of disease spread. Overall, our study provides mechanistic insights on the context in which modular structure influences infectious disease spread in animal populations. We further suggest the use of appropriate null models when data-limited estimates of epidemic consequences in animal groups are necessary.

4.2 Materials and methods

Extended methods are provided in the Supplementary Information.
Figure 4.1: Schematic representation of a (a) homogeneous social network with no modular subdivision, and (b-d) highly modular social networks with (b) high cohesion within subgroups, (c) high network fragmentation, and (d) high subgroup size variation. Each network has 60 nodes. Orange edges indicate interactions within a subgroup and purple edges are between-subgroup interactions. $Q$ is Newman modularity [36] and $Q_{rel}$ is relative modularity.
4.2.1 Measuring modularity

We used modularity ($Q$) proposed by Newman [36] to measure the strength of modular organization in networks. Modularity can be defined as

$$Q = \sum_{k=1}^{K} \left[ \frac{L^w_k}{L} - \left( \frac{L_k}{L} \right)^2 \right] ;$$

where $L_k$ is the total number of edges in a subgroup $k$, of which $L^w_k$ are the edges within the subgroup, and $L$ is the number of total edges in the network. Community structure, or the number and composition of subgroups, for each animal social network was estimated using the Louvain method [149]. The highest possible modularity in a network ($Q_{max}$) is achieved when all individuals in a subgroup $k$ only interact with each other and no edges are present between subgroups (i.e., subgroups are disjointed). In other words, $Q_{max}$ of a network is when $L^w_k = L_k$, and can be written as

$$Q_{max} = \sum_{k=1}^{K} \frac{L_k}{L} (1 - \frac{L_k}{L}) .$$

We measured the relative modularity of networks as $Q_{rel} = \frac{Q}{Q_{max}}$.

4.2.2 Mechanisms of modular organization

To examine the relative contribution of factors such as network size, network fragmentation, average subgroup size, etc., towards network modularity, we ran a mixed effects beta regression model using glmmADMB package (v0.8.3.3) in R (v3.2.3). We treated the species nested within the taxonomic order of the animal group as a random effect in the model to account for any correlation in the network metrics. However, the variance accounted for by the nesting of species was negligible ($\sigma^2 \approx 0$), and we therefore considered only taxonomic order as the random effect in the analysis. We also account for differences between the social organization of different species by defining three broad levels of sociality - relative solitary, social, and fission-fusion; and include sociality as a random effect in the model.
4.2.3 Generation of synthetic modular and null networks

We used the network generator described in [185] to generate networks of a specified modularity and degree sequence while keeping other network properties close to homogeneous null networks. For Fig. 4.3 and 4.4, we generated synthetic modular networks with 10,000 nodes, with an exponential degree distribution with a mean degree of 10. Two types of null networks were generated for each animal social network. *Modular null networks* were created by randomizing within-subgroup connections and between-subgroup connections. The modular null networks therefore possessed identical modular configuration and degree sequence as empirical networks, but were random with respect to other higher-order network properties. We generated *homogeneous null networks* by performing simultaneous edge swaps over within-subgroup and between-subgroup connections, which preserves the local contact heterogeneity but randomizes all higher-order features of the network including network modularity.

4.2.4 Disease simulations

We performed Monte-Carlo simulations of a discrete-time susceptible-infected-recovered (SIR) model of disease spread. As we were interested only in major outbreaks, we considered only those simulations in our calculations where at least 10% of the population acquired infection. Transmission occurred according to a pathogen transmissibility, defined as the probability of transmission from an infected to susceptible host during the period when the host is infectious. The infectious period was assumed to be constant across hosts.
4.3 RESULTS AND DISCUSSION

4.3.1 MODULAR ORGANIZATION AND DISEASE IN ANIMAL SOCIAL NETWORKS

We obtained previously published (and publicly available) social networks for 69 groups across 43 animal species where edges can serve as realistic routes of infection spread. Newman modularity, $Q$, is a commonly used measure to estimate the strength of subdivision in networks [36]. When $Q = 0$, the density of interactions within subgroups is equivalent to density of interactions between subgroups. Higher values of Newman modularity indicate stronger sub-divisions of social networks. Using Newman modularity to compare modular subdivision across networks is, however, problematic as the measure is inherently correlated to the size and number of links in the network [37]. This is because, in principle, we expect the maximum modularity $Q_{\text{max}}$ of all networks to be equal to one. However for smaller real-world networks, $Q_{\text{max}}$ tends to be a much smaller value (Fig. 4.2a). We therefore estimated relative modularity (obtained by normalizing $Q$ with $Q_{\text{max}}$), to facilitate comparison of modular subdivision strength across different animal groups. Animal social networks in our database, ranged from homogeneous ($Q_{\text{rel}} = 0$) to highly subdivided network structure (Fig. 4.2a). Animal interactions are dynamic in nature, and therefore social networks of the same set of individuals can fluctuate between being relatively homogeneous to highly modular over time (Fig. 4.2b).

To gain intuition about how modular organization in animal social networks influences epidemiological outcomes, we first generated homogeneous null networks for each empirical animal social network in our database. In homogeneous null networks, higher-order organization in social networks (such as modular structure, clustering coefficient, cliquishness and degree homophily) are randomized, weakening relative modularity (SI Appendix, Fig. S1), but individual contact heterogeneity (degree
distribution) of the empirical network is preserved. We next performed susceptible-infected-recovered (SIR) disease simulations of a moderately transmissible infectious disease (basic reproduction number, $R_0=1.2$) through these empirical networks and their homogeneous null network counterparts. Based on previous work suggesting a protective effect of modular structure on disease risk [35, 130], we expected a lower disease burden (measured as the proportion of population infected) in empirical networks than in homogeneous null networks. Reduced disease burden was however apparent only in social networks with $Q_{rel} > 0.6$ for moderately spreading infectious disease (Fig. 4.2c, SI Appendix Fig. S2). Additionally, none of the empirical networks demonstrated a major reduction in disease burden for a highly transmissible infectious disease ($R_0 = 4.8$ in SI Appendix, Fig. S3). This implies that the protective effect of modular subdivisions is realized only at high values of relative modularity, but the presence of a threshold depends on pathogen transmissibility (SI Appendix, Fig. S3).

4.3.2 Determinants of modular organization in animal social networks

Fig. 4.2c suggests that the epidemiological consequences of modular structure is context-dependent, depending on the strength of relative modularity, and only above a certain threshold. Additionally, the extent of outbreak size reduction above the modularity threshold was fairly inconsistent across networks. This could be due to varying features of interaction patterns driving modular subdivision in animal social networks. We therefore first turn our attention to some of these network features including the number of individuals present in the social network ($network size$); (log of the) number of subgroups in the social network ($network fragmentation$); preferential association with own subgroup ($subgroup cohesion$, measured as the proportion
Figure 4.2: Animal social networks with modular subdivisions. (a) Values of Newman modularity, $Q$, and maximum possible modularity given the network configuration, $Q_{\text{max}}$, across all animal groups in the database. Point color represents the relative modularity estimated as the ratio of $Q$ over $Q_{\text{max}}$. The dashed line represents $Q = Q_{\text{max}}$ and represents the maximum value of $Q_{\text{rel}} = 1$. (b) Relative modularity values of dynamic interactions in ants (*Camponotus fellah*), raccoons (*Procyon lotor*) and field voles (*Microtus agrestis*), where time points represent consecutive days, weeks, and trapping sessions, respectively. (c) Comparisons between real (solid points) and their homogeneous null networks (tip of arrows) with respect to percentage of infected individuals (outbreak size) due to an outbreak with basic reproduction number, $R_0=1.2$. Point color corresponds to the taxonomic class of the animal group. Social networks with non-significant modular subdivision (as indicated by t-test analysis) have been excluded in panel c. The generated homogeneous null networks preserve only the local heterogeneity of contacts among individuals; the arrows therefore indicate the change in direction and magnitude of outbreak size under the scenario where all higher-order structural complexities (including modular subdivisions) are removed from animal social networks. The shaded area represents the region where empirical networks tend to experience reduced outbreak size (at $Q_{\text{rel}} > 0.6$, SI Appendix Fig. S3). The networks that experienced lower disease burden included social networks of raccoons (*Procyon lotor*), elephant seals (*Mirounga angustirostris*), ants (*Camponotus pennsylvanicus*), bottlenose dolphins (*Tursiops truncatus*), and Australian sleepy lizards (*Tiliqua rugosa*). In panel a and c we used a randomly-selected network snapshot for groups with temporal interaction data.
Table 4.1: Multivariable analysis on determinants of modular organization \( (Q_{rel}) \) across groups of different species. Asterisks and bold text indicate significance. Explanatory variables with VIF > 5 were dropped from the model and therefore their effect sizes were not estimated.

<table>
<thead>
<tr>
<th>Explanatory variable</th>
<th>Effect size</th>
<th>95% confidence intervals</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>-0.42</td>
<td>-0.59 – -0.25</td>
</tr>
<tr>
<td>Network size</td>
<td>-0.07</td>
<td>-0.17 – 0.03</td>
</tr>
<tr>
<td>Network fragmentation*</td>
<td>0.66</td>
<td>0.51 – 0.82</td>
</tr>
<tr>
<td>Subgroup cohesion*</td>
<td>1.14</td>
<td>1.03 – 1.25</td>
</tr>
<tr>
<td>Subgroup cohesion variation</td>
<td>-0.03</td>
<td>-0.14 – 0.08</td>
</tr>
<tr>
<td>Subgroup size average</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Subgroup size variation*</td>
<td>-0.13</td>
<td>-0.26 – -0.01</td>
</tr>
<tr>
<td>Individual degree average</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Individual degree variation</td>
<td>0.07</td>
<td>0.04 – 0.18</td>
</tr>
<tr>
<td>Subgroup degree variation</td>
<td>0.01</td>
<td>-0.08 – 0.10</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Random effects</th>
<th>Variance estimate ( (\sigma^2) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Taxonomic order</td>
<td>0.066</td>
</tr>
<tr>
<td>Sociality</td>
<td>&lt; 0.001</td>
</tr>
</tbody>
</table>

of total contacts that occur within subgroups); variation in subgroup cohesion across the network; average and variation in subgroup sizes; average and variation in number of individual contacts \( (\text{degree}) \); and variation in contacts among subgroups \( (\text{subgroup degree variation}) \). For groups that had multiple temporal snapshots of a social network, we randomly selected one snapshot network for the analysis to avoid intra-group correlation in network metrics (and we analyze these temporal networks separately later on in this section).

The mixed effects beta-regression model identified three network features driving modular organization in animal social networks: subgroup cohesion, network fragmentation and subgroup size variation (Table C.1). Of the three factors, subgroup cohesiveness and network fragmentation had strong positive effects on modular orga-
nization. Cohesiveness can be a direct result of strong social bonds due to similar nutrient requirements (e.g., in spotted hyenas [186]), time spent foraging together or matrilineal descent (e.g., in African elephants [187]). High network fragmentation, on the other hand, can be caused by constraints on resource density, resource distribution or high intra-group competition (e.g., brown spider monkeys [188]). In contrast to network fragmentation and subgroup cohesion, we found variation in subgroup sizes to have a weak negative effect on network modularity. Fluid subgroup sizes are common in many social species (e.g., in spotted hyenas, elks, chimpanzees, bottlenose dolphins, African lions) and can be brought about by changes in resource availability, intra-group aggression, demographic factors (such as sex-ratio), dominance hierarchy, and female reproductive state [186, 188, 189].

Thus far, we have considered only a single snapshot of interactions for each animal group in our database for consistent comparison across species with unequal sampling efforts. However, dynamic interactions between group members in fission-fusion species, species with nomadic individuals, or species with seasonally-driven social interactions can change the level of network modularity over time (Fig.4.2b). We therefore repeated the analysis by considering dynamic interactions for a few animal groups for which such data are available. Consistent with our previous analysis of aggregated interactions, we found animals groups become increasingly modular with higher subgroup cohesion or network fragmentation (Table S1). In addition, variation in subgroup size and variation in contacts among subgroups had a weak negative correlation on relative modularity within animal groups.

4.3.3 The impact of modularity on dynamics of disease spread

We have shown that network fragmentation and cohesion within subgroups are associated with modular structure in animal social networks. Do the two factors also
determine the disease consequences of modular subdivisions in animal social networks? In order to answer this question, we used synthetic modular networks developed using our model proposed in [185]. The model generates networks with tunable strength of modularity while keeping other higher-order network features (such as degree homophily, clustering coefficient, path length) close to the configuration realized in homogeneous networks. Since several structural features of networks can influence disease spread, this approach is appropriate to isolate the epidemiological consequences of modular organization in social networks. Performing SIR disease simulations on synthetic modular networks, we find that the magnitude of outbreak size experienced by low modular networks ($Q_{rel} < 0.6$) is similar to that of homogeneous networks ($Q_{rel} = 0$). Outbreak size is reduced only in extremely modular networks for moderately transmissible pathogens (Fig. 4.3a,b). We corroborate our findings that highly (but not low) modular networks inhibit global parasite transmission by investigating the spread of protozoan parasites *Babesia* in the social networks of field voles [96] (SI Appendix, Fig S11). In addition, we found no evidence of structural trapping of a gastrointestinal parasitic infection in the low modular social network of brown spider monkeys [190] (SI Appendix Fig S12).

We next quantify the robustness of modular social networks towards disease spread in two ways. First, we estimate the minimum level of pathogen contagiousness below which there is no risk of a large outbreak (epidemic threshold, solid line in Fig.4.3a). We find networks with $Q_{rel} > 0.6$ to have a higher epidemic threshold compared to homogeneous and low modular networks, implying that highly modular networks alleviate the risk of large disease outbreaks for low contagious pathogens. Second, for pathogen transmissibility beyond the epidemic threshold, we calculate the minimum level of relative modularity where networks experience at least a 10% reduction in outbreak size as compared to homogeneous ($Q_{rel} = 0$) networks (modularity
threshold, solid line in Fig.4.3b). We find that networks with $Q_{rel} \geq 0.5$, mitigate disease for moderately contagious (transmissibility = 0.08 – 0.16) but not highly contagious pathogens (transmissibility > 0.16). Highly modular networks ($Q_{rel} >0.8$) have a striking impact on disease transmission - outbreak sizes of moderately transmissible pathogens are reduced up to 58% compared to homogeneous networks. Such highly modular networks, although rare, do occur in host interaction networks. In our database three animal species demonstrated $Q_{rel}$ values greater than 0.8 - raccoons ($Procyon lotor$), field voles ($Microtus agrestis$), and northern elephant seals ($Mirounga angustirostris$).

Why doesn’t any degree of modular subdivision reduces disease burden in animal social networks? The relationship between disease burden and modularity depends on the trade-off between local disease transmission within subgroups and global disease spread (Fig.4.3c). Compared to homogeneous networks (with $Q_{rel} = 0$), modular networks experience higher local transmission due to a high density of contacts within subgroups; on the other hand, low inter-subgroup contacts in these networks reduce global disease transmission (Fig.4.3c). This trade-off balances the overall disease spread, and thus modular structure, under most conditions, does not lower disease burden. The conditions under which modularity does reduce disease burden of social networks is largely determined by the two factors underlying network modularity - network fragmentation and subgroup cohesion (Table C.1). Synthetic networks that possess both high subgroup cohesion and high network fragmentation experience lower outbreak size compared to homogeneous and low modular social networks (Fig.4.3c and SI Appendix Fig. S6). In addition, high variation in subgroup sizes along with high subgroup cohesion and fragmentation, further reduce outbreak size and increase outbreak duration in social networks (SI Appendix, Fig. S7). Disease consequences in real animal social networks, however, tend to be driven by subgroup cohesion as they
typically exhibit a much lower range of fragmentation and subgroup size variation as compared to subgroup cohesion (SI Appendix, Fig S5).

4.3.4 MECHANISMS THAT DRIVE THE IMPACT OF HIGH MODULAR STRUCTURE ON DISEASE SPREAD: STRUCTURAL DELAY AND TRAPPING

We now examine the mechanisms by which highly fragmented social networks with cohesive subgroups experience reduced disease burden. The presence of only a few interactions between highly cohesive subgroups increases the amount of time it takes for an infection to spread from one subgroup to another (the structural delay effect, Fig. 4.4a, SI Appendix, Fig S12), causing longer disease outbreaks. When networks are fragmented into highly cohesive subgroups, the structural delay imposed by cohesion causes infection to be localized to a small proportion of subgroups before dying out (Fig. 4.4b). Popularly known as structural trapping [191], this effect also reduces the overall likelihood of a major outbreak by a novel infection in modular social networks. The structural trapping effect (albeit weak) has been observed in a field-study of pneumonia transmission in highly subdivided networks of bighorn lambs (SI Appendix, Fig S13) [192], and others (Table S2). We also find evidences of structural delay in the spread of mycoplasma in the low-modular network but highly cohesive subgroups of house finches described in [193] (SI Appendix, Fig S14). A high probability of extinction associated with the structural trapping effect has been reported in previous theoretical studies of disease spread in spatially-structured metapopulation models [26, 182]. We note that the magnitude of structural trapping and delay depends on the contagiousness of the infection - moderately transmissible pathogens spread slowly through the populations and are therefore able to perceive modular structures present in networks. For rapidly spreading pathogens, highly fragmented
Figure 4.3: Overall disease implications of modular subdivisions. (a) Average outbreak size, measured as the percentage of infected individuals, over increasing subdivided social networks and pathogen transmissibility. Outbreak size values have been normalized to the maximum observed outbreak. The solid line indicates epidemic threshold, viz., the threshold value of pathogen contagiousness below which there is no risk of a large outbreak (> 10% outbreak size, see Materials and Methods). (b) Epidemic robustness of networks with increasing value of relative modularity, measured as the percentage reduction in outbreak size as compared to outbreak size experienced by homogeneous ($Q_{rel} = 0$) networks. The solid line indicates modularity threshold where networks experience at least a 10% reduction in outbreak size. (c) Infection transmission events, expressed as the percentage of total outbreak size, within subgroups (local) and between subgroups (global), pathogen transmissibility = 0.18; (d) Disease implications of modular subdivisions as a function of subgroup cohesion and network fragmentation (measured as the lognumber of subgroups present in the network).
Figure 4.4: Mechanisms behind the effect of modular organization on disease spread of a moderately transmissible pathogen (=0.1). (a) Structural delay effect in a moderately fragmented network (fragmentation = 2.30): we considered the time to disease invasion for a subgroup as the number of time-steps it takes for at least 2% of individuals to acquire infection. Subgroups at the x-axis are sorted according to the increasing disease invasion time. (b) Structural trapping effect in a highly fragmented network (fragmentation = 4.83): high fragmentation and subgroup cohesion localize infection to a small proportion of subgroups in the social network. Structural trapping also increases the likelihood of stochastic extinction of disease outbreaks.

and cohesive networks do not contain but delay the spread of disease to other subgroups, creating only a structural delay effect.

4.3.5 Is knowledge of modular structure necessary (and sufficient) to predict disease outcomes?

Our investigations of disease consequences on synthetic networks suggest that only modular subdivision beyond a certain threshold can reduce disease burden of moderately transmissible pathogens. Other topological features of the network, and in particular heterogeneity in host contact [7], are however bound to confound the effect
of modularity on infectious disease spread. Our work thus far assumes high variation in contacts among individuals [194], although factors including sociality [195] and non-persistent space use [196] are known to reduce variation in individual contact rate. Does then the individual contact heterogeneity affect the way modular organization influences disease spread? We find that for social networks with low contact heterogeneity, high modularity brings about a greater reduction in disease burden compared to networks with high contact heterogeneity (SI Appendix, Fig. S8). These results imply that the local contact patterns should also be considered while estimating the amelioration of disease burden in extremely subdivided social networks.

Although modular subdivision is a common feature observed in many animal groups, it is often correlated with other higher-order network structures, such as clustering coefficient and degree homophily [174]. In empirical networks, it is therefore important to ask whether modular organization has any epidemiological consequence above and beyond these network properties, after accounting for contact heterogeneity. To answer this question we performed disease simulations through 19 animal social networks in our database and their modular null networks. Modular null networks preserve the modularity level and degree distribution of their empirical network counterparts, but randomize all other network structure. We expect a high error rate in outbreak size prediction from modular null networks if other structural features mask the effect of modular subdivisions. We also compared the performance of modular null networks to that of homogeneous null networks where all network structure, except degree distribution, is randomized (similar to the networks used in Fig.4.2c). We find low (but not zero) percentage error in outbreak size predictions from modular null networks (Fig. 4.5, SI Appendix Fig. S9). Homogeneous null networks however perform equally well in predicting disease outcome for most animal social networks, except when relative modularity of empirical networks is high ($Q_{rel} >$
0.6). We note that this $Q_{rel}$ value is similar to the threshold identified above (Fig. 4.2c, SI Appendix Fig. S3). For moderately transmissible infections, animal social networks experience low outbreak size at this level of modular structure due to the structural trapping effect. Homogeneous null networks do not preserve the modular structure of the empirical networks, and therefore overestimate the true disease burden. In contrast to moderately spreading infectious diseases, slow ($R_0 < 1$) and rapidly spreading infections are unaffected by higher-order complexities of the networks, and therefore outbreak size estimates of homogeneous and modular null networks are similar across all levels of relative modularity (SI Appendix, Fig. S10). To confirm the results of disease simulations, we compared the performance of modular and homogenous null networks to predict outbreak size of two Salmonella enterica strains in the Australian sleepy lizards, Tiliqua rugosa, using data described in [197] (SI Appendix, Fig. S15). Congruent to the theoretical predictions, homogeneous and modular network produced identical and accurate outbreak size predictions for the low transmissible strain. For the moderately transmissible strain, on the other hand, modular null networks performed better in estimating the true disease burden.

4.4 Conclusions

Although the disease implications of group living in animal societies has been much debated, there has been equivocal evidence of the role of modular structure on alleviating disease burden of group living species. More importantly, the features of interaction patterns that cause animal social networks to be subdivided have not been identified before. In this study, we resolve the ambiguity of past research on disease consequences of network modularity in three ways. First, we introduced a normalized version of Newman modularity called relative modularity ($Q_{rel}$) to allow compar-
Figure 4.5: Percentage error in outbreak size predictions using modular and homogeneous null networks for 19 animal social networks due to an outbreak with $R_0=1.2$. Percentage error is calculated as $(S_{emp} - S_{null})/S_{emp} \times 100$, where $S$= outbreak size, $emp$ = empirical network, $null$ are modular or homogeneous null networks of the empirical network. The social networks are ordered according to the increasing value of relative modularity (red solid curve, secondary y-axis). The shaded region indicates the range of percentage error values below 15%. Animal social network abbreviations: MT = Macaca tonkeana, MM = Macaca mulatta, BB = Bison bison, CF = Camponotus fellah, DC = Dairy cattle, DR = Desmodus rotundus, BA = Brachyteles arachnoides, CCR = Crocuta crocuta, MF = Macaca fuscata, CC = Cercopithecus campbelli, TT = Tursiops truncatus, PC = Papio cynocephalus, HM = Haemorhous mexicanus, TR = Tiliqua rugosa, CP = Camponotus pennsylvanicus, MA = Mirounga angustirostris. Numbers denote separate groups of the same species.
ison across networks of different sizes, and identified two distinct mechanisms that can lead to modular structure. Second, we systematically studied the epidemiological consequences of modular organization by using a model [185] that allows for the mechanistic generation of synthetic modular networks and null networks (of empirical animal social networks) while minimizing the presence of higher-order network properties. Third, we combined published social networks across 43 animal species with this theoretical network model and biologically realistic disease simulation to derive insights on disease consequences of mechanisms of modular organization in animal societies.

It has been hypothesized that subgroups in animal social networks reduce the disease costs of group living [35, 179]. Contrary to this hypothesis, our study shows that disease burden is largely unaffected by modular subdivisions in animal groups except when social networks are extremely fragmented and have high cohesion within subgroups. High subgroup cohesion and fragmentation in networks structurally trap moderately transmissible infections to a few subgroups and therefore reduce the overall disease burden. We note that although overall disease burden is unaffected at most levels of modularity, the design of control strategy will vary for different levels of network modularity. An effective intervention should aim to minimize global spread of infection for low network modularity, but reduce local transmission at high network modularity. Since our analysis is based on a systematic investigation of contagion spread in synthetic modular networks, the results of this study are potentially applicable across a wide range of systems, including spreading processes at higher ecological scales (Table S2), infectious disease spread in human interaction networks and transportation networks, perturbation spread in metabolic and ecological networks, as well as spread of information through communication networks.
Since this study involved meta-analysis of social networks across a broad range of taxonomic groups, we made a number of simplifying assumptions that should be considered before extending the results to a specific system. First, we filtered the weights assigned to social interactions between individuals to construct unweighted networks. This is because the impact of weights (whether they represent frequency, duration or intensity of contact) on transmission potential is generally unclear and usually context-dependent. We advise leveraging transmission studies in captive populations or historical disease spread data, where available, to identify the appropriate weighting criterion of contact networks. Second, we assume the aggregated static networks and temporal "snapshots" of dynamic networks in our database coincide with the temporal scale relevant to infection spread in our disease simulations. Consideration of the relevant time interval of animal interactions relative to the transmission mode and infectious period of the pathogen is, however, crucial to developing accurate network models of disease spread [26, 198]. As many animal interactions can be infrequent or intermittent, particularly between subgroups, aggregating interactions over a small time-window may ignore these fleeting contacts and over-estimate the modularity of social networks. On the other hand, pooling interactions over large time windows may amplify the role of temporary contacts and consequently the social subgroups might appear more connected. In addition, chronic infections with long infectious period will be relatively unaffected by high network modularity if the time-scale of network fluctuation is shorter than the average infectious period. Acute infections with short infectious period, however, will be highly sensitive to the state of network modularity at the time of disease outbreaks.

This study suggests the presence of a high modularity threshold above which social networks experience reduced outbreak size. However, we caution against the use of single modularity threshold that is applicable to all systems. Multiple behavioral,
ecological and epidemiological factors including pathogen infectiousness (Fig. 4.3b), infectious period, local contact heterogeneity, seasonality of contacts, and transmission coupled with recruitment of susceptible through births can influence the $Q_{rel}$ threshold and the extent to which network fragmentation and subgroup cohesiveness structurally trap and delay infection spread. Therefore, to assess the disease implications of modular structure observed in specific social groups, we recommend the use of null networks, as presented in this study. Comparison of disease consequences on empirical network data to those on a hierarchy of null networks clarifies whether modular structure is important. If natural history and limited observations suggest a species’ network will be highly modular, then models that incorporate modularity should be used; otherwise, homogeneous null models (based on basic knowledge about network size and local heterogeneity) may be sufficient when data-limited estimates of epidemic consequences are necessary.
Chapter 5

Disease implications of animal social network structure: a synthesis across social systems

5.1 Introduction

Host social behavior plays an important role in the spread of infectious diseases. Socially complex species from honeybees to African elephants live in large groups and are considered to have elevated costs of pathogen transmission due to high contact rates [38, 39]. Previous studies have tested hypotheses about the disease costs of sociality by associating group size with infection transmission [199, 200]. Beyond a simple dependence on group size, however, recent work in the field of network epidemiology has shown that infectious disease spread largely depends on the organization of infection-spreading interactions between individuals [40, 41, 42, 201]. Indeed, even when interactions between individuals are assumed to be homogeneous, the expectation of higher disease costs of group living has been mixed [199, 200, 202].

Mathematically, social networks describe patterns of social connections between a set of individuals (nodes) by placing edges connecting the nodes [194, 203, 204]. The advantage of social network analysis is that it integrates heterogeneity in interaction patterns at individual, local and population scales to model global level processes, including the spread of social behavior and infectious diseases [194, 205, 206, 207]. A fundamental individual-level characteristic relevant to the spread of social or biological contagion in networks is the number of direct social partners, associates or contacts,
capturing the interaction necessary for transmission. In recent years, network analysis tools have allowed for rapid advances in our understanding of how individual interaction rates are related to the risk of acquiring infection [12, 59]. While much attention has been focused on the implications of individual sociality, the disease implications of a species’ social system remains unclear.

By quantifying group-level metrics that describe global structures in interaction patterns, the network approach provides a unique opportunity to examine the disease costs of species social system. The role of higher-order network structures such as degree heterogeneity (Fig. 5.1A), group cohesion (Fig. 5.1D), network fragmentation (Fig. 5.1E), and global clustering coefficient (Fig. 5.1F) on infectious disease spread is complex, but is relatively well understood (see network structure definitions in Table S1)[4, 46, 208]. For example, as degree heterogeneity (or variation in the number of social partners) in a network increases, the epidemic threshold (i.e., the minimum pathogen transmissibility that can cause large outbreaks) approaches zero. However, the probability of epidemic outbreaks is lower in networks with high degree variance for moderately and highly transmissible pathogens. [4]. Network clustering and modularity measures the tendency of nodes to form cliques and subgroups, respectively (Fig. 5.1). Although the dynamics of infectious disease spread remain largely unaffected in networks with moderate levels of clustering and modularity, extreme levels of these metrics in networks reduce epidemic size and prolong epidemic outbreaks [46, 208].

Recent mathematical models predict that the network structure of socially complex species can serve as a primary defense mechanism against infectious disease by lowering the risk of disease invasion and spread [209]. It remains uncertain, however, whether the structure of social networks naturally observed in less-complex social systems mediates infectious disease risk and transmission. A systematic examina-
tion of the disease costs associated with species social system requires a comparative approach that isolates unique structural characteristics of social connections, while controlling for population size, data collection methodology and type of interaction recorded. However, comparing networks across different taxonomic groups has proven to be a difficult task, with only a few cross-species network comparisons previously published in the literature [44, 45, 46].

In this study, we conduct a quantitative comparative analysis across 47 species to investigate whether social network organization alone, without the presence of physiological or behavioral immune responses, can reduce the disease costs of group living for complex and less-complex social systems. This is achieved in three steps. First, we use phylogenetically-controlled Bayesian generalized linear mixed models to identify social network structures which are predictive of a social system. Second, we perform computational experiments of infection spread to evaluate the role of various global network metrics in curbing the invasion of a novel pathogen and limiting the spread of infectious diseases. Finally, we compare three key epidemiological outcomes (epidemic probability, epidemic duration and epidemic size) among three types of social systems. We hypothesize that a social species can mitigate disease costs associated with group living through the organization of their social structure. However, we expect the presence of alternate disease defense mechanisms to also play an important role: social insects, for example, use social immunity as a primary strategy to minimize disease transmission; the structure of the social network in such species may not be effective in preventing future outbreaks or reducing disease transmission. Our analysis, by broadening the scope of network analysis from species-specific analysis to a meta-analytic approach, offers new perspective on how social structure strategies mediate the disease costs of group living. A better understanding of the association between network structure and different social systems can facilitate investigations
on other evolutionary and ecological hypotheses on group living, social complexity, communication, population robustness and resilience to extrinsic population stressors.

5.2 Materials and methods

5.2.1 Dataset

We first conducted electronic searches in Google Scholar and popular data repositories, including Dryad Digital Repository and figshare for relevant network datasets associated with peer-reviewed publications. We used the following terms to perform our search: "social network", "social structure", "contact network", "interaction network", "network behavior", "animal network", "behavior heterogeneity" and "social organization". Only studies on non-human species were considered in our primary search. Network studies not reporting interactions (such as biological networks, food-web networks) were excluded. By reviewing the quality (i.e., whether enough information was provided to accurately reconstruct networks) of published networks datasets, we selected 666 social networks spanning 47 animal species and 18 taxonomic orders. Edge connections in these networks represented several types of interactions between individuals, including dominance, grooming, physical contact, spatial proximity, direct food-sharing (i.e. trophallaxis), foraging, and interactions based on the asynchronous use of a shared resource. Fig. 5.2 summarizes the species, the number of networks and the reported interaction types contributed by each taxonomic order represented in the study.

5.2.2 Classifying species' social system

Developing a definition of social structure that encompasses the continuum of social systems across diverse taxonomic groups is challenging. Consequently, we followed
Figure 5.1: A stylized illustration of the global network measures used (in the final model) to identify the structural differences in the social networks among different social systems. (a) Degree heterogeneity, measured as the coefficient of variation (CV) in the frequency distribution of the number of social partners (known as the degree distribution). Shown is the degree distribution of a homogeneous network (CV ≪ 1), and an exponential degree distribution of a network with large variation in individual degrees (CV = 1). (b) Degree homophily (ρ), or the tendency of social partners to have a similar degree. Shown is an example of a disassortative network, wherein high degree individuals tend to associate with low degree individuals (ρ < 0), and assortative degree networks, where high degree individuals tend to form social bonds with each other (ρ > 0). (c) Global betweenness centrality, that measures the tendency of nodes to occupy central position within the social network. Shown is an example of a network with low global betweenness centrality and a network with high global betweenness centrality. Node colors represent the betweenness centrality values - nodes with darker colors occupy more central positions within the network. (d) Group cohesion measures the tendency of individuals to interact with members of own group. The network to the left has three low cohesive social groups, while the network to the right has highly cohesive social groups where most of the interactions occur within (rather than between) groups. (e) Network fragmentation, measured as the log-number of the social groups (modules) present within the largest connected component of a social network. Shown is an example of low (left) and highly (right) fragmented network. (f) The global clustering coefficient measures the ratio of (closed) triplets in versus the total number of all possible triplets in the networks. (g) Network diameter is the longest of all shortest paths between pairs of nodes in a network. Shown is an example of a network with low network diameter (longest of shortest paths = 3) and a similar network with network diameter of 5, indicated by red colored edges.
[210] and [211] to classify species into three broad categories of social structure based on the degree of association between adults during activities such as foraging, traveling, sleeping/resting and rearing offspring. Relatively solitary species were defined by infrequent aggregation or association between adults outside of the breeding period, and lack of synchronized movements in space by adults. Examples of relatively solitary species in the database include the desert tortoise (Gopherus agassizii), wild raccoons (Procyon lotor), and the Australian sleepy lizard (Tiliqua rugosa). Recent studies suggest that the social structure of a species traditionally considered as solitary can be complex [181, 212]. We therefore categorized the three species as relatively solitary and not solitary. Species that aggregate for one or more activities, but have unstable or temporally varying group composition were classified as gregarious. Examples of gregarious species in our database include bottlenose dolphins (Tursiops truncatus), bison (Bison bison), Indiana bats (Myotis sodalis), female Asian elephants (Elephas maximus), sociable weavers (Philetairus socius), golden-crowned sparrows (Zonotrichia atricapilla) and guppies (Poecilia reticulata). Species characterized by a permanent or temporary stable social hierarchy were classified as socially hierarchical. Examples of socially hierarchical species include carpenter ants (Camponotus fellah), yellow baboons (Papio cynocephalus), male elephant seals (Mirounga angustirostris) and spotted hyenas (Crocuta crocuta). We note that animal social behavior is being increasingly recognized to span a continuum from solitary to eusocial [213, 214, 215], with most species showing some level of fission-fusion dynamics [216]. The division of social systems into three discrete, albeit arbitrary, categories allows for simple distinctions in the organization of network structure and disease risks among species that are characterized by different complexity in group living behavior.
Figure 5.2: Phylogenetic distribution of animal species represented in the social network dataset used in this study. Numbers next to the inner ring denote the total networks available for the particular species. The inner and the middle ring is color coded according to the taxonomic class and the social system of the species. The colors in the outer ring indicates the type of interaction represented in the network, and whether the interactions were coded as (direct) interactions or association in our analyses (in brackets). The tree was constructed in the Interactive Tree Of Life (http://itol.embl.de/) from the NCBI taxonomy database (http://www.ncbi.nlm.nih.gov/Taxonomy/).
5.2.3 Identifying unique network structures of species’ social system

To examine the structure of social networks associated with our three classified social systems, we used a Bayesian generalized linear mixed model (GLMM) approach using the MCMCglmm package in R [217], with the species’ social system as the response (categorical response with three levels - relatively solitary, gregarious and socially hierarchical). The following network measures were included as predictors in the model (see Table S1 in Supporting information for definitions and Fig.5.1 for illustrations): degree heterogeneity, degree homophily, global clustering coefficient, weighted clustering coefficient, transitivity, average betweenness centrality, weighted betweenness centrality, average group size, network fragmentation, group cohesion, relative modularity and network diameter. Network fragmentation (i.e., the number of groups within the largest connected component of the social network) and Newman modularity was estimated using the Louvain method [149]. Relative modularity was then calculated by normalizing Newman modularity with the maximum modularity that can be realized in the given social network [46, 185]. The rest of the network metrics were computed using the Networkx package in Python (https://networkx.github.io/).

We controlled for network size and density by including the number of nodes and edges as predictors, and average edge weight was included to control for data sampling design. To control for phylogenetic relationships between species, a correlation matrix derived from a phylogeny was included as a random factor. The phylogenetic relationship between species was estimated based on NCBI taxonomy using phyloT (http://phylot.biobyte.de). We controlled for repeated measurements within groups, animal species, the type of interaction recorded, spatial scale of recorded networks, and edge weighting criteria by including group, taxa, interaction type (association vs. interaction) and edge weight type (weighted vs. unweighted) as random effects in the
analysis. As the spatial scale of data collection can influence network structure (Table S3, Supporting information), we specified sampling scale (social sampling vs. spatial sampling) as random effect in all our analyses. Studies that collected data on specific social groups were categorized as social sampling, and those that sampled all animals within a fixed spatial boundary were labelled as spatial sampling.

All continuous fixed-effects were centered (by subtracting their averages) and scaled to unit variances (by dividing by their standard deviation) to assign each continuous predictor with the same prior importance in the analysis [83]. Since network measures can be highly correlated to each other, variance inflation factor (VIF) was estimated for each covariate in the fitted model, and covariates with VIF greater than 5 were removed to avoid multicollinearity. We used a weakly informative Gelman prior for fixed effects and parameter-expanded priors for the random effects to improve mixing and decrease the autocorrelation among iterations [218]. Specifically, a $\chi^2$ distribution with 1 degree of freedom was used as suggested by [219]. We ran three MCMC chains for 15 million iterations, with a thinning interval of 1000 after burn-in of 50,000. Convergence of chains was assessed using the Gelman-Rubin diagnostic statistic [220] in the coda package [221].

Groups of certain species in our database were represented with multiple networks, each summarizing a set of interactions occurring in a discrete time period. To ensure that such animal groups were not over-represented in the original analysis, we performed a cross-validation of our analysis by random sub-sampling. Specifically, we repeated the analysis 100 times with a random subset of the data composed of (randomly selected) single networks of each unique animal group in our database. An average of coefficient estimates across the multiple subsamples was then calculated and compared to the coefficients estimated using the full dataset.
Evaluating the role of weak ties in driving structural differences in species’ social system

The analysis described in the previous section assumes equal importance of all edges recorded in a social network. To examine the role of weak ties in driving the structural differences between the three social systems, we removed edges with weights lower than a specified threshold. Four edge weight thresholds were examined in detail: 5%, 10%, 15% and 20%. Specifically, all edges with weights below the specified threshold were removed to obtain filtered social networks. Next, the phylogenetically-controlled Bayesian mixed model analysis described in the previous section was repeated to determine the structural difference between the filtered networks of the three social systems. We ran four separate models, each with one of the four thresholds.

5.2.4 Disease implications of network structure and species’ social system

We considered disease costs of the three social systems with synthetic experiments based on a computational disease model, and followed up with statistical analysis of the results.

Disease simulations

We performed Monte-Carlo simulations of a discrete-time susceptible-infected-recovered (SIR) model of infection spread through each network in our database. For disease simulations, we ignored the weights assigned to social interactions between individuals, because the impact of interaction weight (whether they represent contact duration, frequency or intensity) on infection spread is generally not well understood epidemiologically. Transmissibility of the simulated pathogen was defined as the
probability of infection transmission from an infected to susceptible host during the infectious period of the host. Assuming infection transmission to be a Poisson process and a constant recovery probability, the pathogen transmissibility can be calculated as $T = \frac{\beta}{\beta + \gamma}$, where $\beta$ and $\gamma$ is the infection and recovery probability parameter, respectively [7]. Each disease simulation was initiated by infecting a randomly chosen individual in the social network. At subsequent time steps every infected individual in the network could either transmit infection to a susceptible neighbor with probability parameter $\beta$ or recover with probability $\gamma$. The disease simulations were terminated when there were no remaining infected individuals in the network. We performed disease simulations with a wide range of transmissibility values (0.05 to 0.45, with increments of 0.05), by varying infection probability ($\beta$) and assuming a constant recovery probability ($\gamma = 0.2$ or average infectious period of 5 days). To investigate the effects of recovery probability on the behavior of pathogen spread, we repeated disease simulations with a similar range of transmissibility values as before (0.05 to 0.45), but with a longer infectious period (10 days or $\gamma = 0.1$). For each combination of pathogen transmissibility and social network, 500 simulations of disease spread were carried out and summarized using three measures: (a) epidemic probability, the likelihood of an infectious disease invasion turning into a large epidemic (outbreaks that infect at least 15% of the population) (b) epidemic duration, the time to epidemic extinction, and (c) epidemic size, the average percentage of individuals infected in an epidemic outbreak.

EVALUATING DISEASE OUTCOMES OF NETWORK STRUCTURE AND SPECIES’ SOCIAL SYSTEM

Three separate Bayesian GLMMs, one corresponding to each outbreak measure (epidemic probability, epidemic duration, and epidemic size), were fit to establish disease
costs of network measures associated with species’ social system. To evaluate the role of network structure on the probability of large outbreaks, pathogen transmissibility and network measures included in the final model of the previous analysis were included as predictors (Table 5.1). We repeated the analysis with the species’ social system as predictor to directly estimate the vulnerability of different social structure towards disease transmission.

In all models, the effective number of nodes (i.e., the number of individuals with degree greater than zero), network density and the size of the largest connected component of the network were also included as controlling predictors. As before, we controlled for the presence of phylogenetic correlations, group identification, animal species, interaction type, edge weight type, and sampling scale of networks. Minimally informative priors were used for fixed effects (normal prior) and (co)variance components (inverse Wishart; [217]). We ran three MCMC chains for 100 thousand iterations, with a thinning interval of 10 after burn-in of 2000, and assessed convergence using the Gelman-Rubin diagnostic statistic [220] in the coda package. To make posthoc comparisons within the models, we performed pairwise comparisons between the three social systems with a Tukey adjustment of \( P \) values, using the lsmeans R package [222].
Table 5.1: Effect size estimates of the Bayesian generalized linear mixed models examining the characteristics of social network structure among the three social systems: relatively solitary, gregarious and socially hierarchical. Shown are the posterior means of the expected change in log-odds of being in focal social system (column headers), as compared to the base social system (row headers), with one-unit increase in the network measure. The 95% credible intervals (i.e., the coefficients have a posterior probability of 0.95 to lie within these intervals) are included in brackets. Significant terms with pMCMC < 0.05 are indicated in bold, where pMCMC is the proportion of MCMC samples that cross zero.

<table>
<thead>
<tr>
<th></th>
<th>Focal Base</th>
<th>Relatively solitary</th>
<th>Gregarious</th>
<th>Socially hierarchical</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Degree heterogeneity</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Relatively solitary</td>
<td></td>
<td>-3.96 [-7.57, -0.33]</td>
<td>-9.46 [-15.21, -3.87]</td>
<td>-6.39 [-11.67, -1.34]</td>
</tr>
<tr>
<td>Gregarious</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Socially hierarchical</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Degree homophily</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Relatively solitary</td>
<td></td>
<td>-0.18 [-1.66, 1.17]</td>
<td>-1.69 [-3.80, 0.25]</td>
<td>-1.64 [-3.25, 0.09]</td>
</tr>
<tr>
<td>Gregarious</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Socially hierarchical</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Global betweenness centrality</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Relatively solitary</td>
<td></td>
<td>0.68 [-2.31, 3.76]</td>
<td>0.36 [-2.91, 3.82]</td>
<td>0.27 [-2.56, 2.12]</td>
</tr>
<tr>
<td>Gregarious</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Socially hierarchical</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Global clustering coefficient</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Relatively solitary</td>
<td></td>
<td>-0.06 [-2.49, 2.47]</td>
<td>-3.40 [-6.56, -0.24]</td>
<td>-3.30 [-5.82, -0.88]</td>
</tr>
<tr>
<td>Gregarious</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Socially hierarchical</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Group cohesion</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Relatively solitary</td>
<td></td>
<td>-0.60 [-2.98, 1.84]</td>
<td>-0.40 [-3.23, 2.12]</td>
<td>0.97 [-1.14, 3.05]</td>
</tr>
<tr>
<td>Gregarious</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Socially hierarchical</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Network fragmentation</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Relatively solitary</td>
<td></td>
<td>3.94 [0.74, 7.26]</td>
<td>0.11 [-4.01, 4.12]</td>
<td>-3.27 [-6.11, -0.51]</td>
</tr>
<tr>
<td>Gregarious</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Socially hierarchical</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Network diameter</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Relatively solitary</td>
<td></td>
<td>-1.79 [-5.00, 1.45]</td>
<td>1.46 [-2.79, 5.52]</td>
<td>2.86 [-0.31, 5.89]</td>
</tr>
<tr>
<td>Gregarious</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Socially hierarchical</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

96
5.3 Results

5.3.1 Unique network structures associated with species’ social system

The final model (after removing collinear predictors) consisted of seven global network measures - degree heterogeneity, degree homophily, global betweenness centrality, global clustering coefficient, group cohesion, network fragmentation and network diameter (Fig. 5.1, Table 5.1). Out of the five random effects included in the model (phylogeny, group identification, interaction type, edge type, sampling scale), phylogeny explained a large portion of the variance (Table S2, Supporting information), indicating that there is a substantial phylogenetic correlation within the social systems. Of the three social systems (relatively solitary, gregarious and socially hierarchical), the social networks of relatively solitary species demonstrated the largest variation in the number of social partners, or degree heterogeneity (Table 5.1). In contrast, socially hierarchical species had the least variation in number of social partners, and experienced a local social environment that is not as well inter-connected; this is evident by the low global clustering coefficient of their social networks as compared to other social systems (global clustering coefficient, Table 5.1). In terms of network fragmentation (which was calculated on the largest connected component of networks), the social networks of gregarious species were the most subdivided into socially cohesive groups. No statistically significant differences were observed between the social systems with respect to other network metrics. Table S3 of Supporting information reports the average coefficient estimates of all seven global network metrics from the cross-validation analysis; all estimates were within the 95% credible interval of the effect sizes reported in the full model (Table 5.1). We also find that the organization of social networks depends on the sampling scale of social associations, but not
on the type of interactions recorded. For example, networks measured at a population scale rather for social groups tended to have low local connectivity, as measured by the global clustering coefficient, and low global betweenness centrality (Table S4, Supporting information).

5.3.2 Disease costs of network structure and species’ social system

Our previous analysis revealed that only a few features of social networks are significant in distinguishing the three social systems. Next we ask: Do these key topological differences mediate differential disease costs of each social system? To answer this question, we first examined how degree heterogeneity, clustering coefficient and network fragmentation influence epidemic risk and transmission of low, moderate and highly transmissible pathogens (Fig. 5.3; see Fig. S2, S4 in Supporting information for results on an extended range of pathogen transmissibility values and Fig. S5 for results on disease simulations with extended infectious period). High variation in individual sociality (i.e., high degree heterogeneity) in social networks was predictive of small and short epidemic outbreaks for low transmissible pathogens. Moderately spreading pathogens in network with high degree heterogeneity led to less frequent, shorter epidemics that infected a smaller proportion of the population (degree heterogeneity, Fig. 5.3). The presence of cliques in social networks was associated with prolonged but small outbreaks of low transmissible pathogens, and higher epidemic risk of moderately transmissible infections (global clustering coefficient, Fig. 5.3). Subdivisions of networks into socially cohesive groups (high fragmentation) was associated with reduced risk of lowly transmissible infections becoming large epidemics; outbreaks that did reach epidemic proportion were shorter and infected a lower proportion of the population. Conversely, highly contagious pathogens caused frequent, large, and
Figure 5.3: Role of network structures in influencing disease transmission summarized as epidemic probability (likelihood of large outbreaks infecting at least 15% of individuals in the network), average epidemic duration (time to epidemic extinction), and average epidemic size (percent of individuals infected in the social network), for low (=0.05), moderate (=0.15) and highly (=0.45) transmissible pathogens. The three global network measures shown are the ones that were found to differ among the three social systems (Table 5.1). DH, degree heterogeneity; CC, global clustering coefficient; NF, network fragmentation. Error bars represent 95% credible intervals. Credible intervals that do not include zero suggest significant association with disease transmission (red = significant effect, black = effect not significant).
prolonged epidemic outbreaks in networks with high network fragmentation (network fragmentation, Fig. 3).

Consequently, socially hierarchical species experienced elevated risk of epidemic outbreaks of moderately transmissible pathogen due to homogeneous individual connectivity (low degree heterogeneity) and high global connectivity (low network fragmentation) nature of their social networks (epidemic probability, Fig. 5.4, Fig. S3 and S5 in Supporting information). The highly fragmented networks of gregarious species were more vulnerable to frequent, large, and prolonged epidemic outbreaks of highly transmissible pathogens as compared to other social systems. Given that degree heterogeneity and network fragmentation is associated with shorter outbreaks of low transmissible pathogens (Fig. 5.3, Fig. S3 and S6 in Supporting information), epidemic duration of less transmissible pathogens was lowest in gregarious species, followed by relatively solitary species (epidemic duration, Fig. 5.4, Fig. S3 and S6 in Supporting information). For moderately contagious pathogens, highly fragmented networks of gregarious species experienced longer epidemic outbreaks as compared to relatively solitary and socially hierarchical species.

Role of weak ties in distinguishing species’ social system, and disease implications

When the weakest 5% edges were removed from all weighted networks, the structural differences between the three social systems were observed mainly in two network metrics - degree heterogeneity and network fragmentation. Similar to the unfiltered networks (Table 5.1), the 5% filtered social networks of relatively solitary species demonstrated the highest variation in number of social partners; and 5% filtered networks of gregarious species were more fragmented compared to relatively solitary and socially hierarchical species (Table S5, Supporting information). When the
Figure 5.4: Disease costs of social systems due to social network structure. Disease cost has been quantified in terms of epidemic probability, average epidemic duration and average epidemic size for low (=0.05), moderate (=0.15) and highly (=0.45) transmissible pathogens. Error bars represent standard errors, and different letters above the bars denote a significant difference between the means (P < 0.05).

Weakest 10% and 15% edges were removed, the global network measures across all social systems were similar to each other, except for one important difference. Both 10% and 15% filtered networks of social species (gregarious and socially hierarchical) demonstrated a statistically significant higher global betweenness centrality, or higher global connectivity than relatively solitary species (Table S6, S7 and S8, Supporting information). Disease simulations through 20% edge weight filtered social networks revealed no differences in epidemiological outcomes between the three social systems for all except low pathogen transmissibility (Fig. S7, Supporting information). For slow spreading pathogens, networks of relatively solitary species experienced larger epidemic sizes as compared to social species.

5.4 Discussion

It is becoming increasingly clear that the impact of an infectious disease on a population depends on the organization of infection-spreading interactions between individu-
uals rather than group size. [40, 41, 42, 46]. Since organization of social network structure concurrently impacts the transmission of information and infectious diseases, it has critical implications for understanding the evolutionary tradeoffs between social behavior and disease dynamics. The disease implications of social network structure can differ depending on the evolutionary trajectory of social systems. For instance, social complexity can emerge as a result of selective pressures of past infectious diseases, and therefore may have the ability to lower the risk of transmission of future infectious diseases [209]. Conversely, the patterns of social interactions may not provide protection from disease transmission in species that use alternate defense mechanisms (physiological or behavioral) to combat disease spread once it is introduced in the population [29, 223, 224]. In this study, we assessed whether network structure alone (in absence of physiological or behavioral disease defense mechanisms) can reduce the risk of infectious disease transmission in different social systems, using comparative methods on an extensive database of animal social networks.

Our analysis compares global structural features associated with social networks of species classified into three social systems: relatively solitary, gregarious and socially hierarchical. The evidence that we present here suggests that, at the least, relatively solitary, gregarious, and higher social organizations can be distinguished from each other based on (i) degree of variation among social partners (i.e. degree heterogeneity), (ii) local connectivity, as indicated by the presence of cliques within the social networks (i.e, global clustering coefficient), and (iii) the extent to which the social network is divided into cohesive social groups (i.e., network fragmentation). Remarkably, the structure differences between social systems exist in spite of differences in data collection methods and type of interaction recorded.

Social species are typically assumed to have a skewed degree distribution (for e.g. bottlenose dolphins [164], wire-tailed manakins [225]), which implies that a small
proportion of individuals have a large number of social partners. Our results, however, show that degree heterogeneity in relatively solitary species can be much higher than social species. Large variation in the number of social connections in relatively solitary species may simply arise due to a high variation in spatial behavior as compared to social species [181, 196]. A homogeneous degree distribution in socially hierarchical species, such as ants and savanna baboons, could allow for efficient and equitable information transfer to all individuals [226, 227]. Low global clustering coefficient, as observed in socially hierarchical species, is known to increase network resilience and stability in response to perturbations such as the removal or death of individuals [194].

Our results also show that social networks of gregarious species are the most subdivided (but not disconnected) into cohesive social groups. The presence of many but small, socially cohesive groups within social networks of gregarious species can be explained based on the behavioral tendency to switch affiliative partners; as a result, individuals form consistent social bonds with a only small subset of individuals [32]. Many gregarious species also form groups based on sex or age class, kinship and functional roles [228], and modular subdivisions has been shown in theoretical models to promote behavioral diversity and cooperation [229, 230]. Gregarious species may therefore limit the size of their social subgroups to maximize benefits of cooperation, making their social networks subdivided [231].

Our results show that the observed structural differences between the three social systems are primarily driven by the presence weak ties in their social networks. The reason why filtering out weak weighted edges removes most structural differences between social systems lies in their organization of weak ties. Individuals of social species disproportionately allocate effort among their social connections in order to maintain overall group connectivity (Fig. S1, Supporting information) and are also
known to have high social fluidity [232]. Removing weak ties from networks of social species therefore increases variation in individual connectivity (degree heterogeneity), with a relatively minor decrease in their global connectivity (average betweenness centrality). Consequently, the global connectivity of social species in 10%-15% filtered networks is significantly higher than relatively solitary species.

Previous studies have typically focused on group size as the key parameter impacting disease transmission and group living costs. However, the expectation of higher disease costs of group living has yielded mixed results [199, 200, 202], which can be explained in part by the presence of group-level behavioral [224, 233] and physiological defense [234] against infection spread, as well as the presence of chronic social stress [235, 236]. While group size might be easy parameter to measure, it does not capture the complex spatio-temporal dynamics of most animal societies. By performing disease simulations over empirical networks with different interaction types, we consider a range of infectious diseases with different modes of transmission, including infections that spread via direct contact, via the environment and via indirect (vector) interactions. Our analysis shows that the organization of social patterns may not provide general protection against pathogens of a range of transmission potential. We note that our results on epidemic size and duration are specific to pathogens that follow SIR (susceptible-infected-recovered) infection dynamics. The outcome of epidemic probability, however, is expected to be similar across different models of infectious disease spread (such as infections that provide temporary immunity or chronic infections).

We find that socially hierarchical species experience longer outbreaks of low transmissibility infections and frequent epidemics of moderately contagious infections because of low variation in individual and local connectivity (i.e., degree heterogeneity and global clustering coefficient) as compared to other social systems. Networks with
low degree heterogeneity are known to experience steady protracted outbreaks, in contrast to explosive rapid outbreaks fueled by superspreaders in high degree heterogeneity networks [4, 7, 237]. High global clustering coefficient is also believed to create redundant paths between individuals making it harder for slow spreading infections to encounter new susceptible individuals and percolate throughout the network, prolonging infection spread [174].

In our disease simulations, highly fragmented social networks of gregarious species experienced frequent epidemics of highly contagious infections, and longer epidemics of moderately to highly transmissible pathogens. Our recent work has shown that infection spread in highly fragmented networks gets localized within socially cohesive subgroups (structural trapping), which enhances local transmission but causes structural delay of global infection spread [46]. In addition, our results suggest that highly transmissible pathogens are able to avoid stochastic extinction in fragmented networks by reaching "bridge" nodes, but experience delay in transmission due to the presence of structural bottlenecks.

As this study involved comparisons of social networks across a broad range of taxonomic groups and data sampling methods, we made a number of assumptions that could shape the results. First, because the impact of edge weights on disease transmission can be context-dependent, depending on the type of interaction, transmission mode of pathogen, and the relative time scale of network collection and pathogen spread, we have chosen to not include edge weights while performing our computational disease experiments. Future meta-analytic studies can leverage a growing number of transmission studies to explicitly incorporate the role of contact intensity on disease transmission [238, 239]. Second, we assume that social contacts remain unaltered after an infection is introduced in population. Presence of infection, however, can alter the social connectivity of hosts [240]. Future species specific studies can
take advantage of host specific experimental manipulations, where possible, to gain in-depth insight towards host behavior - infection feedback [43, 206]. Finally, in our network database there were some systematic differences in data-collection methodologies across social systems. Specifically, all data for relative solitary species were collected by sampling individuals over a specified spatial range, because definition of social groups for these species can be vague. As observations of direct interactions in relatively solitary species is rare, all networks of relatively solitary species in our database were based on direct or indirect spatial associations. Although the meta-analysis described in this study controlled for such biases in data-collection, the results should be interpreted as a conceptual understanding about the differences between the social systems in terms of empirical networks that have been published in the literature, and not as a general prediction about the differences in social systems.

Overall, our results suggests that the organization of social networks in gregarious species are more efficient in preventing outbreaks of moderately contagious pathogens than socially hierarchical species. Conversely, networks of socially hierarchical species experience fewer outbreaks of fast spreading infectious diseases as compared to gregarious species. The question of why this is so warrants detailed future investigations of the eco-evolutionary trajectory of social connectivity in the two social systems. It is likely that the organization of social networks in socially hierarchical species may have evolved to prevent outbreaks of highly transmissible pathogens, while relying on alternate group-level disease defense mechanisms (including sanitary behaviors, allogrooming, and the use of antimicrobials) to prevent outbreaks of low to moderate transmissibility infections. Since the social networks included in the meta-analysis were selected regardless of the presence of infectious diseases in the populations, the organization of network structure could also reflect the selection pressure of past infections, or the presence of other evolutionary drivers [241].
5.5 Challenges and opportunities

The sociality of animal species has been traditionally classified based on qualitative phenotypes and life history traits, and the classification typically differs between taxonomic groups. While this categorization scheme is convenient, it does not capture the continuum of social behavior. As a step forward, recent studies have proposed quantitative indices of sociality [214, 242]. The results of our study support the potential use of network structure as a means of quantifying social complexity across taxonomic boundaries. Similar predictive meta-analyses can also be used to identify species that are outliers in the current sociality classification system based on the organization of their social structure.

However, we need to overcome several challenges before robust comparative analysis can be performed on social networks across broad taxonomic groups to address such issues. First, comparing network structure across taxonomic groups where data is aggregated over different spatio-temporal scales is challenging. Aggregating interactions over small time-periods may omit important transient interactions, whereas aggregating data over long time-periods may lead to a saturated network where distinguishing social organization may be difficult. Spatial constraints and environmental heterogeneity can also impose a considerable influence on the social network structure [96, 243]. Additionally, the consideration of relative time scale of animal interaction and infectious period of pathogen is critical in making accurate predictions of disease spread. Future comparative studies should therefore consider standardizing data over time and spatial scales.

The second challenge lies in defining a social group that is consistent across species and social systems. This is difficult because the general rules of defining groups in social species (including definitions of spatial and temporal coherence) do not hold
true for relatively solitary and fission-fusion species. Network size correlates to sampling intensity in many cases, and is therefore a poor proxy to group size.

The third challenge for comparative studies of animal social networks is in effectively controlling for inherent biases in data collection methodologies across taxonomic groups. As direct observation of interactions is difficult in relatively solitary species, social networks are usually constructed based on direct or indirect spatial associations (rather than interactions) between individuals in a population (rather than a local group). Social network studies of relatively solitary species are also relatively sparse compared to social species.

In conclusion, we note that there is enormous potential of adopting a comparative approach to study the commonalities and differences in social networks across a wide range of taxonomic groups and social systems. Future studies can use this approach to quantitatively test several evolutionary and ecological hypotheses, including the ones on group living, social complexity, information transfer, and resilience to population stressors. To facilitate in-depth meta-analyses of network data, we encourage researchers to accompany animal network datasets with the following details: data sampling method, location of the data collection, type of population monitored (captive, semi-captive, free-ranging), edge definition, edge weighting criteria, node attributes (such as demography), temporal resolution of data, temporal and spatial aggregation of the data, proportion of animals sampled in the area, and population density. When exact measurements of these data attributes are difficult, using reasonable approximations or proxies would be more useful than no information.
Chapter 6

REVEALING TRANSMISSION MECHANISMS OF INFECTIOUS DISEASES THROUGH
EMPirical CONTACT NETWORKS

6.1 Introduction

Host contacts, whether direct or indirect, play a fundamental role in the spread of infectious disease through host populations [46]. Traditional epidemiological models, however make simplistic assumptions such as homogeneous mixing of individuals, no social structure, which are often unrealistic and in many cases yield unreliable predictions of disease spread [7]. Therefore, in recent years, network approaches to modeling infection spread have gained popularity because they explicitly incorporate host interactions that mediate infection spread. Formally, in a contact network model, individuals are represented as nodes, and an edge between two nodes represents an interaction that has the potential to transmit the infection. An exact contact network model requires (i) knowledge of the transmission mechanism of the spreading pathogen (required to define an edge in contact network), (ii) sampling of all individuals in a population, and (iii) all disease-causing interactions between the individuals. In addition, the accuracy of disease predictions depends on the knowledge of epidemiological characteristics of the spreading infection, including the rate of transmission given a disease-causing interaction between two individuals and the rate of recovery of infected individuals. The use of modern technology in recent years, including RFID,
GPS, radio tags, proximity logger and automated video tracking has enabled the collection of detailed movement and contact data, making network modeling feasible.

Despite the technology, logistical and financial constraints still prevent data-collection of each individual and each interaction occurring in a host population [244, 245]. More importantly, limited knowledge about the pathogen makes it challenging to identify the mode of infection transmission, define disease causing contacts between individuals, and measure the probability of transmission given contact [239]. Laboratory techniques of unraveling transmission mechanisms usually take years to resolve [246]. Defining accurate contact networks underlying infection transmission in human infectious disease has been far from trivial [7]. For animal infectious disease, limited knowledge about host behavior along with the characteristics of the spreading pathogen makes it particularly difficult to define an exact contact network, which has severely limited the scope of network modeling in animal and wildlife epidemiology [247].

Lack of knowledge about host-pathogen characteristics has prompted the use of several indirect approaches to identify the link between network structure of animal societies and disease spread. A popular approach has been to explore the relation between social network position on disease risk [13, 40, 248]. Another approach is to use proxy networks, such as movement networks, and network based on spatial proximity or home-range overlap, as a surrogate to all direct and indirect contacts between individuals [195, 249, 250]. A recent approach, called the k-test procedure, explores a direct association between infectious disease spread and contact network by comparing the number of infectious contacts of infected cases to that of uninfected cases [201]. However, several challenges still remain in identifying the underlying contact networks of infection spread in a host population. First, it is often unclear how edge weights (whether duration, frequency or intensity) relate to transmission
risk unless validated by transmission-experiments [238]. Furthermore, the relevance of low-weighted edges in a contact network is ambiguous [251]. The interaction network of any social group will appear as a fully connected network if monitored for a long period of time. As fully-connected contact networks rarely reflect the transmission of an infection through a population, one may ask whether low weighted links can be ignored, and what constitutes an edge-weight threshold below which interactions are epidemiologically irrelevant? Second, many previous approaches ignore the dynamic nature of host contacts. The temporal dynamic of contacts between hosts, that is building and dissolving of contacts over time, however is crucial in determining the order in which contacts occur, which regulates the spread of an infectious disease through the host population [252]. Therefore, static contact networks often fail to approximate disease spread through dynamic host networks [253, 254]. Third, an approach that establishes the predictive power of a network to describe the dynamics of disease spread is lacking. Although spatial proximity or home-range overlap as proxies for infectious contact may be a reasonable assumption to make considering the limited data that is available for certain wildlife systems, such contact network models may be uninformative. This is because, by mixing disease-spreading contacts with those that are epidemiologically irrelevant, the disease predictions from such networks may not be reliable or accurate. Finally, to our knowledge, none of the previous approaches allow testing of competing hypotheses about disease transmission mechanisms which (may) generate distinct contact patterns and consequently different contact network models.

All of these challenges demand an approach that can allow hypothesis testing between different contact network models while taking into account the dynamics of animal interactions and missing data due to incomplete network sampling. In this study, we introduce a computational tool called INoDS (Identifying Network models
of infectious Disease Spread) that provides statistical evidence for the underlying contact network of disease transmission in host populations. Our tool can infer contact network models of a wide range of infectious disease types (SI, SIS, and SIR) and can be easily extended for complex disease spread models. We develop a three-step approach to enable power analysis and hypothesis testing on contact network models of infectious disease spread (Figure. 6.1). Our tool provides inference on dynamic and static contact network models, and is robust to common forms of missing data. Using two real-world datasets, we highlight the two-fold application of our approach – (i) to identify the specific contact network associated with a particular host population, and (ii) to gain epidemiological insights (including the transmission mechanism, role of the quality of host contacts) of a host-pathogen system that can be leveraged to construct contact networks for predictive modeling of disease spread in host populations with similar ecological conditions at other geographical locations.

6.2 Results

The primary purpose of INoDS is to assess statistical evidence towards the contact network that is hypothesized to be relevant for the spread of an infection through a host population. INoDS also provides epidemiological insights of the spreading pathogen by estimating the rate of pathogen transmission through the edges of the contact network, and facilitates hypothesis testing on its mode of transmission. The inference follows a three step procedure. First, the tool estimates the unknown social (\( \beta \)) and asocial (\( \alpha \)) transmission parameter. The social transmission parameter quantifies those infection acquisition events that are explained by the edge connections of the network hypothesis, while the asocial transmission parameter captures the infection events that are unexplained by the network hypothesis.
INoDS utilizes the observed infection time-series and dynamic (or static) interaction data to provide statistical evidence towards contact network hypotheses (or hypothesis) using a three step procedure. Inferential steps: In the first step, the tool estimates two parameters for a contact network hypothesis - transmission rate $\beta$, which represents the component of infection transmission that is contributed by the network connections, and $\alpha$ that quantifies the component of infection propagation unexplained by the contact network. Second, the significance of $\beta$ is estimated by comparing the magnitude of social and asocial transmission rate for all observed infection events. The predictive power of the network hypothesis is evaluated by comparing the likelihood of the infection time-series data given the network hypothesis to an ensemble likelihoods derived by permuted null networks. Third, the marginal likelihood for the given contact network is calculated, which is then used to perform model comparison (using Bayes Factor, BF) in cases where multiple contact network hypotheses are available.
In the second step, the likelihood of the observed infection time-series data given the network hypothesis is compared to the likelihood distribution of an ensemble of null networks (generated by permuting the edge connections of the dynamic network, while controlling for the number of nodes and edges present in the network). The network hypothesis is considered to have a high predictive power if its likelihood is higher than the null likelihoods at 5% significance level. In the third step, marginal (Bayesian) evidence is calculated for the contact network hypotheses; Bayesian evidence can then be used to perform model selection where more than one contact network hypotheses exists.

In the sections that follow, we evaluate the accuracy of the tool in recovering the transmission parameters and its ability to identify the underlying contact network under common types of missing data (missing individuals, missing contacts and missing infection cases). We further demonstrate the application of INoDS by using two empirical dataset: (i) spread of an intestinal pathogen in bumble bee colonies, and (ii) salmonella spread in Australian sleepy lizards.

6.2.1 INoDS performance

We evaluated the performance of INoDS in model discrimination and recovering the unknown transmission parameters. To do so, we generated hypothetical infection time-series data by performing computer simulations of pathogen spread on a synthetic dynamic network. The synthetic network was generated by first establishing a random network (generated by the configuration model [119]) at time-step $t = 0$. For each of the following time stamp, we permuted 10% of randomly chosen edges using a double-edge swap procedure [119], for a total of 100 time-steps.
Figure 6.2: INoDS performance on recovering the social (\( \beta \)) and the asocial (\( \alpha \)) transmission parameter of simulated susceptible-infected (SI) model of infectious disease spread. Disease simulations were performed on dynamic networks with 100 nodes, Poisson degree distribution, mean degree of 4, and 100 discrete time snapshots. Each boxplot summarizes the results of 10 independent disease simulations; the horizontal line in the middle is the mean of estimated parameter values, the top and the bottom horizontal line is the standard deviation, and the tip of the vertical line represents the maximum/minimum value. The solid red line represents one-to-one correspondence between the \( \beta \) value estimated by INoDS and the value of \( \beta \) used for disease simulations. Since the simulations were performed on a known synthetic network, the value of asocial transmission parameter is expected to be zero.
To test the ability of INoDS to precisely estimate the social transmission parameter $\beta$, we performed 10 independent simulations of SI disease spread with transmission rates ($\beta$) ranging from 0.01 to 0.1. Model accuracy was determined by comparing the estimated value of transmission parameter $\beta$ to the value of $\beta$ used to perform disease simulations. Since the synthetic network dataset did not contain any missing data, model accuracy was also tested by evaluating the deviation of the estimated asocial transmission parameter $\alpha$, from the expected value of zero. In Figure 6.2, we show that INoDS is able to recover the true value of $\beta$ and $\alpha$, and that the accuracy of INoDS is independent to the spreading rate of the pathogen. Inclusion of the asocial transmission parameter ($\alpha$) significantly improves the parameter estimation accuracy of $\beta$ when either the network data or disease surveillance is incomplete (Appendix Figure 1). The estimate of the transmission parameter is therefore fairly accurate even under missing data conditions, although parameter estimates are most sensitive to the proportion of missing nodes in the observed network, followed by missing edges (Appendix Figure 2).

Robustness of INoDS and previous approaches to missing data

The performance of INoDS was evaluated in terms of sensitivity and specificity in distinguishing the true contact network model from an ensemble of null networks with randomized edge connections. The performance was tested against three potential sources of error in data-collection – sampling a subset of individuals in a population (missing nodes), incomplete sampling of interactions between individuals (missing edges), and infrequent health diagnosis of individuals (missing cases). Specifically, we randomly removed 10-80% of nodes, edges or infection cases from the simulated
dataset described above. Sensitivity refers to the proportion of times the observed (true) network was correctly classified as statistically significant ($P < 0.05$, corrected for multiple comparison). Specificity was defined as the proportion of times a network with the same degree distribution as the observed network but randomized edge connections was classified as statistically insignificant.

We find that INoDS correctly assigns the observed network with a higher statistical power across a wide range of missing data scenarios (Figure 6.3). The sensitivity and specificity of the tool remains high under imperfect disease surveillance and is unaffected by increasing percentage of unobserved infection cases. For data with incomplete sampling of nodes, sensitivity of INoDS remains close to one when at least 50% of nodes are observed in the network. We also find that the performance of the tool is unaffected by the edge density of the underlying contact network.

Next, we compared the performance of INoDS to two previous approaches – $k$-test procedure and network position test (Figure 6.3, Supplementary information). The $k$-test procedure involves estimating the mean infected degree (i.e., number of direct infected contacts) of each infected individual in the network, called the $k$-statistic. The $P$-value in $k$-test is calculated by comparing the observed $k$-statistic to a distributed of null $k$-statistics which is generated by randomizing the node-labels of infection cases in the network [201]. Network position test compares the degree of infected individuals to that of uninfected individuals. The observed network is considered to be epidemiologically relevant when the difference in average degree between infected and uninfected individuals exceeds (at 5% significance level) the degree difference in an ensemble of permuted networks. Figure 6.3 (extended figure) summarizes the results. We find that $k$-test and network position test were sensitive to all types of missing data. Sensitivity of $k$-test procedure rapidly declined with increasing percentage of missing nodes, edges of the observed contact network, and infection cases. Out of the
three approaches, network position test proved to be the least reliable as its sensitivity was close to zero even when no data were missing.

Since the $k$–test procedure and network position test have been primarily used in the context of non-dynamic networks, we repeated this analysis on a simulated dataset where disease simulations were performed on a static network. Appendix Figure 5 demonstrates that even for observed network that are not dynamic, INoDS outperforms the $k$-test procedure and network position test in correctly identifying the true contact network and rejecting false randomized networks (Appendix Figure 5).

6.2.2 Applications to empirical data-sets

We demonstrate the application of INoDS to perform hypothesis testing on contact networks of infectious disease spread, identify transmission mechanisms and infer
transmission rate using two real-world datasets. The first dataset is derived from the study by [12] that examined the spread of an intestinal pathogen (*Crithidia bombi*) within the colonies of social bumble bee *Bombus impatiens*. The second dataset examines the spread of *Salmonella enterica* within two wild populations of Australian sleepy lizards *Tiliqua rugosa*.

**Determining transmission mechanism and role of contact intensity: case study of the spread of *Crithidia bombi* in bumble bee colonies**

The data consists of dynamic networks of bee colonies (N = 5-7 individuals), where edges represent direct physical contacts that were recorded using a color-based video tracking software. A bumble bee colony consists of a single queen bee and infertile workers. Infection experiments were performed on seven colonies; in five colonies infection was introduced by their naturally infected queens, while in the remaining two colonies one forager was randomly selected to be artificially infected. Infection progression through the colonies was tracked by regularly screening the faeces of each bee in the colony, and the infection timing was determined using the knowledge of the rate of replication of *C. bombi* within its host intestine. [12] showed that the infection timings of susceptible bees in the colony was associated with their frequency of contacts with infected nest-mates rather than the duration of contacts. Here, we extend the results of the previous study by addressing three specific questions: (1) Do networks where edges represent physical contact have a high predictive power to predict the spread of *C. bombi* than random networks? (2) Should the edges in contact network be weighted (by frequency or duration) of contacts, or unweighted?, and (3) do weak ties between individuals contribute to infection transfer?

Here, we analyzed the spread of *Crithidia* in the two artificially infected colonies (colony UN1 and UN2). To test whether direct contacts are predictive of *Crithidia*
spread, we constructed dynamic contact networks for each colony where edges represent close proximity between individuals (<1cm, as described in [12]). Edges in the dynamic network were updated every day until the experiment was terminated. To test the hypothesis about the role of edge weights, two types of dynamic contact networks were constructed for each colony - networks where edges were weighted with respect to frequency of contacts, and networks accounting for the duration of contacts. We note that the dynamic contact networks constructed were fully connected (i.e., all individuals were connected to each other in the network) at all time steps. As fully connected network rarely describe the dynamics of infection spread, we sequentially removed weak ties in the weighted contacts networks to obtain edge filtered networks. Specifically, edges with weights less than 10^-50% of the highest edge weight were removed to generate contact network hypotheses with a range of edge-densities. Corresponding to the two types of edge filtered weighted networks (frequency and duration weighted), two types of unweighted (binary) contact networks were also constructed by removing edge weights on the filtered weighted contact networks.

Figure 6.4 shows the parameter estimates for the four contact network hypotheses at different edge weight thresholds. For duration (weighted and binary) networks, infection transfer could not be explained by social connections when a high percentage of weak edge weights were removed from the contact networks (20% for binary networks and 15% for weighted networks; faded bars in Figure 6.4 A and C). Among the two types of edge weights, frequency weighted contact networks (with 5% of weakest edge weights removed) was associated with highest Bayesian evidence in both the analyzed bee colonies (Figure 6.4D, Appendix Figure E.6D), although in colony UN2, binary networks with 15% of weakest edge weights removed had equivalently high Bayesian evidence (Figure 6.4B).
Figure 6.4: Identifying the contact network model of *Crithidia* spread in bumble bee colony (colony UN2). Edges in the contact network models represent physical interaction between the bees. Since the networks were fully connected, a series of filtered contact networks were constructed by removing weak weighted edges in the network. The x-axis represents the edge-weight threshold that was used to remove weak edges in the network. Two types of edge weights were tested - (a) duration and (b) frequency of contacts. In addition, across all ranges of percent weak edges removed, the two types of weighted network were converted to (c-d) binary networks. The results shown are estimated values of social transmission parameter $\beta$, and estimated values of asocial transmission parameter $\alpha$, for the different contact network hypotheses. The faded bars correspond to networks where the asocial force was higher than the social force for all transmission events (which indicates that the network hypothesis does not explain the spread of *Crithidia* infection). Numbers above bars indicate the log Bayesian (marginal) evidence of the networks that were detected to have statistically significant higher predictive predictive power as compared to an ensemble of null networks ($P < 0.05$, corrected for multiple comparisons).
Identifying transmission mechanism with imperfect disease data: case study of *Salmonella enterica* spread in Australian sleepy lizard populations

This dataset monitors the spread of the commensal bacterium *Salmonella enterica* in two separate wild populations of Australian sleepy lizard *Tiliqua rugosa*. The two sites consisted of 43 and 44 individuals respectively, and these represented the vast majority of all resident individuals at the two sites (i.e., no other individuals were encountered during the study period). Individuals were fitted with GPS loggers and their locations were recorded every 10 minutes for 70 days. Salmonella infections were monitored at the two populations using cloacal swabs on each animal once every 14 days. Consequently, the disease data in this system collected at fortnightly intervals do not identify the onset of each individual’s infection timing. We used a SIS (susceptible-infected-susceptible) disease model, assuming that sleepy lizards recover from salmonella infections, and eventually get reinfected.

We tested the hypothesis that salmonella spreads in the Australia sleepy lizards through spatial proximity of individuals, a hypothesis previously explored in [197]. Proximity networks were constructed by assuming a contact between individuals whenever the location of two lizards was recorded to be within 14m distance of each other. The dynamic networks at both sites consisted of 70 static snapshots, with each snapshot summarizing a day of interactions between the lizards.

Employing Bayesian data augmentation method, we assumed the actual infection timings to be unobserved parameters, which were sampled (along with the two transmission parameters) for each contact network hypothesis using the data on timing of diagnoses for salmonella infections. We found that the likelihood of salmonella infection spreading through the weighted spatial proximity network was significantly
greater than the null expectation at both the sites (Figure 6.5). As compared to unweighted networks, networks with edges weighted by frequency of contacts had higher marginal (Bayesian) evidence at both the sites, indicating that the occurrence of repeated contacts between two spatially proximate individuals, rather just the presence of contact between individuals is important for Salmonella transmission.

6.3 Discussion

Network modeling of infectious disease spread is becoming an increasingly popular approach, as the quality of data that can be collected from animal populations has dramatically improved in recent years owing to the use of modern technology. However, the concepts of power analysis and hypothesis testing are still underdeveloped in network modeling, even though such approaches are widely recognized as key elements to establish the level of 'informativeness' and appropriateness of a model [255, 256]. Our ability to define a contact network relies on our knowledge about the host behavior, and the dominant mode of transmission of the pathogen. Since such knowledge is either derived from expert knowledge (which can be subjective) or laboratory experiments (which are time- and resource-intensive), it is essential to conduct an a priori analysis of contact network models to avoid uninformative or misleading disease predictions.

In this study we therefore present INoDS as a tool that performs network model selection and establishes the predictive power of a contact network model to describe the observed infection spread. INoDS also provides epidemiological insights about the spreading pathogen by enabling hypothesis testing on different transmission mechanisms, and estimating the rate of pathogen transmission through contacts between individuals (social transmission parameter, $\beta$). Unlike previous approaches, INoDS
Site Transmission mechanism $\beta$ $\alpha$ Predictive power Evidence
1 Direct transmission, binary 0.011 0.034 0.03 -1019.33
1 Direct transmission, weighted* 0.376 0.037 0.03 -219.44
2 Direct transmission, binary 0.061 0.042 0.15 -775.80
2 Direct transmission, weighted* 2.734 0.075 <0.001 -250.61

**Figure 6.5: Identifying transmission mechanisms of Salmonella spread in Australian sleepy lizards.** Dynamic network of proximity interactions for a total duration of 70 days between (a) 43 lizards at site 1, and (b) 44 lizards at site 2. Each temporal slice summarizes interaction within a day. Edges indicate that the pair of individuals were within 14m distance of each other (referred as physical interactions), and the edge weights are proportion to the frequency of physical interactions between the node pair. Green nodes are the animals that were diagnosed to be not infected at that time-point, red are the animals that were diagnosis to be infected and grey nodes are the individuals with unknown infection status at the time-point. We hypothesized that the proximity networks could potentially explain the observed spread of *Salmonella* in the population. The results are summarized as a table. Bold numbers indicate that the network hypothesis was found to have statistically significant high predictive power as compared to an ensemble of null networks. The network hypothesis with the highest log Bayesian (marginal) evidence at each site is marked with an asterisk (*).
is robust to missing network data, imperfect disease surveillance, and can provide network inference for a range of disease spread models. The tool can thus be used to provide inference on contact networks for a variety of infection spread occurring both in human and non-human species. Inferring the role of dynamic contacts on infectious disease spread requires the knowledge of either order or timing of infection of individuals in the network. In practice, constraints on data collection (for e.g., due to infrequent health assessments), or infection diagnostics (for e.g., due to sub-clinical infection, poor diagnostics) precludes the knowledge of precise timing of infection acquisition in a host population. To overcome this challenge, our tools assumes the infection timings in a host population to be unobserved, and uses data on infection diagnosis instead to provide inference on contact networks.

As such, our approach addresses a growing subfield in network epidemiological theory that uses statistical tools to infer contact network models using all available host and disease data [244, 257]. The proposed tool in this study can be used to tackle several fundamental challenges in the field of infectious disease modeling [258]. First, INoDS can be used to perform model selection on contact network models that quantify different transmission modes; this approach therefore facilitates the identification of infection-transmitting contacts and does not rely on laboratory experimentation (or subjective expert knowledge). Second, INoDS can be used to establish the predictive power of proxy measures of contact in cases where limited interaction data is available. For example, spatial proximity, home-range overlap or asynchronous refuge use are commonly used as a proxy of contact in wild animal populations [59, 181, 248]. INoDS establishes the epidemiological significance of such assumptions by comparing the likelihood of infection spread occurring along the edges of the proxy contact network to the likelihoods generated from an ensemble of random networks. Third, it is well known that not all contacts between the hosts have the same potential for
infection transfer. The heterogeneity of host contacts in a network model is typically captured through edge weights, but it often not clear which type of edge weights (frequency, duration or intensity) is relevant in the context of a specific host pathogen system [251]. Through model selection of contact networks with similar edge connection but different edge weight criterion, INoDS can help establish a link between edge weight and the risk of transmission across an edge in the contact network.

We demonstrate the application of INoDS using two real world datasets. In the first dataset, we used INoDS to determine the role of edge weight and edge weight type on the predictive power of the contact network. To accurately model the spread of *Crithidia* gut protozoan in bumble bee colonies, we show that the contact networks weighted with respect to frequency, rather than the duration, of contacts between individual have higher predictive power with respect to observed patterns of transmission. Our results therefore support the original finding of the study [12], where individual’s risk of infection was found to be correlated with their contact rate with infected nest-mates. However, our analysis extends the previous findings by comparing the observed patterns of transmission against null expectations from random networks, and assessing the likelihood of competing network hypotheses. We find weak ties below a certain threshold do not play an important role in infection transfer. Contact networks where such weak weighted edges have been removed, therefore, demonstrate higher predictive power than fully connected networks. In the next empirical example, we explore two transmission mechanisms of a commensal bacterium in wild population of Australian sleepy lizards - direct transfer of bacterium through host physical proximity and indirected transfer measured by the extent of host’s home-range overlap. Our results show that contact networks of host’s spatial proximity predict the infection spread in lizard population rather than contact network based on home-range overlap. These findings support a previous study which suggests that
the bacterial transmission could occur between closely located animals rather than indirect environmental transmission [197].

The current version of INoDS, assumes the infection has no latent period, and that the infectiousness of infected hosts and susceptibility of naive hosts is equal for all individuals in the population. In future, these assumptions can be easily relaxed to incorporate more complex formulations of pathogen spread through a host population. For instance, heterogeneity in infectiousness of infected hosts and susceptibility of naive hosts can be incorporated as random effects in the model by assuming the two to follow a Gaussian distribution. Disease latency can be incorporated using the data-augmentation technique described in the paper.

To summarize, we have designed a simple and general framework that provides inference on contact network underlying infectious disease spread, given the host behavior and infection incidence data. Our approach is robust to missing data, and does not require information on the actual infection timings in the population, which is rarely available. The tool described in the this study, on one hand, can be used to establish the power of a contact network model to make reliable disease predictions; on the other hand, the tool can be used to gain epidemiological insights (such as the mode of infection transfer, role of quality of host contacts) for host-pathogen systems. Since data-collection for network analysis can be labor-intensive and time-consuming, it is essential to make decisions on how limited data collection resources are deployed. Based on the sensitivity analysis of our tool to missing data, we learn that the data-collection efforts should aim to sample as many individuals in the population as possible, since missing nodes have the greatest impact (rather than missing edges) on the predictive power of network models. In future, INoDS can be used to improve targeted disease management and control by identifying high-
risk behaviors and super-spreaders of a novel pathogen without relying on expensive transmission experiments that take years to resolve.

6.4 Materials and methods

Here we describe INoDS, a computational tool for identifying underlying contact networks of infectious disease spread. INoDS provides evidence towards a contact network model, and enables discrimination of competing contact network hypotheses, including those based on pathogen transmission mode, edge weight criteria and data collection techniques. In order to run INoDS, the following two types of data are required as an input: infection time series data, which includes infection diagnoses (coded as 0 = not infected and 1 = infected), and timestep of diagnosis for all available nodes in the networks; and edge-list of dynamic (or static) contact network. An edge-list format is simply the list of node pairs (each node pair represents an edge of the network), along with the weight assigned to the edge and time-step of interaction (optional), with one node pair per line. The tool can also be used on unweighted networks - an edge weight of one is assigned to all edges in this case. The time-step input is not required when the analysis is to be performed on a static network. The software is implemented in Python, is platform independent, and is freely available at (https://github.com/prathasah/INoDS-model).

6.4.1 INoDS formulation

We assume that at each instance the potential to become infected for an individual \( i \) depends on two processes - (a) social transmission \( \beta \), that is captured by the edge connections in the contact network hypothesis, and (b) asocial transmission \( \alpha \), that
represents those infection events that are not explained by the edge connections. The
infection receiving potential, \( \lambda_i(t) \), of individual \( i \) at time \( t \) is then calculated as:

\[
\lambda_i(t) = 1 - \exp\{-\beta k_i(t-1) - \alpha\}
\]  

(6.1)

where \( k_i(t-1) \) denotes the number of infected connections of the focal individual \( i \) at the previous time-step \( (t-1) \) as determined by the network hypothesis, and the two transmission parameter are \( > 0 \).

The log-likelihood for all observed timings of infection in a population given the
network hypothesis \( (H_A) \) can therefore be estimated as:

\[
\log(D|H_A, \beta, \alpha) = \sum_n \log[\lambda_n(t_n)] + \sum^t \left( \sum^m \log[1 - \lambda_m(t)] \right)
\]  

(6.2)

where \( t_n \) is the time of infection of individual \( n \). The first part of equation 6.2
therefore estimates the log likelihood of all observed infection acquisition events. The
second part of the equation represents the log-likelihood of susceptible individuals \( m \)
remaining uninfected at time \( t \).

6.4.2 Data augmentation for unknown infection and recovery time

Calculation of network likelihood using equation 6.2 requires the knowledge of exact
timing of infection, \( t_1, ...t_n \), for \( n \) infected individuals in the population. However in
many cases, the only data that is available are the timings of when individuals in a
populations were diagnosed to be infected, \( d_1, ...d_n \). We therefore employ a Bayesian
data augmentation approach to estimate the actual infection timings in the disease
dataset [259]. Since the actual infection time \( t_i \) for an individual \( i \) is unobserved,
we only know that the time of infection for individual \( i \) lies between the interval

129
\((L_i, d_i]\), where \(L_i\) is the last negative diagnosis of individual \(i\) before infection acquisition. Within this interval, the individual could have potentially acquired infection at any time-step where it was in contact with other infected individuals in the network. Assuming incubation period to be one time-step, we can therefore represent the potential set of infection time as \(t_i \in \{g_i(t_i - 1) > 0, L_i < t_i \leq d_i\}\), where \(g_i(t_i - 1)\) is the degree (number of contacts) of individual \(i\) at time \(t_i - 1\). The data augmentation proceeds in two steps. In the first step, the missing infection times are imputed conditional on the possible set of infection times. In the next step the posterior distributions of the unknown parameters are sampled based on the imputed data. We performed data imputation using inverse transform sampling method, which is a technique of drawing random samples from any probability distribution given its cumulative distribution function [260]. For infections that follow a SIS or SIR disease model, it is essential to impute the recovery time of infected individuals for accurate estimation of infected degree. To do so, we adopt a similar data augmentation approach as described before. We performed data imputation by drawing random samples (using uniform priors) from the set of possible recovery time-points using the inverse transform sampling technique.

6.4.3 Estimating the transmission parameters

We adopt a Bayesian approach to estimate the social \(\beta\) and asocial \(\alpha\) transmission parameters. We use Markov chain Monte Carlo (MCMC) technique with flat priors to obtain, after a burn-in period, the joint posterior density for the parameters. Using Bayes’ Theorem, the joint posterior distribution for a set \(\Theta\) of parameters can be written as
\[ P(\Theta|D, H) = \frac{\mathcal{L}(D|H, \Theta)P(\Theta|H)}{\mathcal{E}(D|H)} \propto \mathcal{L}(D|H, \Theta)P(\Theta|H) \]  \hspace{1cm} (6.3)  

where \( D \) is the infection time-series data, \( H \) is the contact network hypothesis, and \( P, \mathcal{L}, \mathcal{P}, \mathcal{E} \) are the shorthands for the posterior, the likelihood, the prior and the evidence, respectively. In all our results, we use a uniform prior on \([0, 1000]\) for each transmission parameter. Posterior distributions were obtained using the \textit{emcee} package implemented in Python [261].

**Interpreting the asocial transmission parameter**

In principle, inclusion of the asocial transmission parameter in eq. 6.1 is similar to the asocial learning rate used in the network based diffusion analysis approach in the behavior learning literature [262, 263]. However, in contrast to the asocial learning rate (which quantifies the rate of spontaneous learning), \( \alpha \) in INoDS formulation serves to improve the estimation of the pathogen transmissibility parameter, \( \beta \), when either the contact network or infection spread is not completely sampled (Appendix Figure 2). The magnitude of \( \alpha \) can be used to (approximately) estimate the amount of missing data. The percentage transmission events where the asocial force (\( = \alpha \)) is greater than the the social force (\( = \beta.k \)) proportionately increases with higher percentage of missing network data, but is less sensitive to missing information on infection cases (Figure 3, Supplementary information).

**6.4.4 Testing the predictive power of a contact network hypothesis**

To assess the significance of the contact network hypothesis, we compare the likelihood of the infection time-series data given the contact network and estimated transmission parameter (i.e., \( \mathcal{L}(D|H, \Theta) \)) to a distribution of likelihoods of infection data for
randomized networks. Randomized networks are generated by randomizing edge connections in the contact network hypothesis, which preserves the edge-density in the permuted networks. A $p$-value is calculated as the proportion of randomizations which generate a likelihood more extreme than the likelihood under the network hypothesis. A contact network with a $p$-value of less than 0.05 is considered to demonstrate a substantive greater predictive power than the null expectation.

6.4.5 Performing model selection of competing network hypotheses

To facilitate model selection in cases where there are more than one network hypothesis, we compute marginal likelihood of the infection data given each contact network model. The marginal likelihood, also called the Bayesian evidence, measures the overall model fit, i.e., to what extent the infection time-series data can be simulated by a network hypothesis ($H_1$). Bayesian evidence is based on the average model fit, and calculated by integrating the model fit over the entire parameter space:

$$P(D|H) = \int \mathcal{P}(\Theta|H) \mathcal{L}(D|H, \Theta) d\Theta$$  \hspace{1cm} (6.4)

Since it is difficult to integrate Eq.6.4 numerically, we estimate the marginal likelihood of network models using thermodynamic integration, or path sampling [264] method implemented in _emcee_ package in Python.

Model selection can be then performed by computing pair-wise Bayes factor, i.e. the ratio of the marginal likelihoods of two network hypotheses. The log Bayes’ factor to assess the performance of model $M_1$ over model $M_2$, is expressed as:

$$\log(B_{21}) = \log(P(D|H_2)) - \log(P(D|H_1))$$  \hspace{1cm} (6.5)
The contact network with a higher marginal likelihood is considered to be more plausible, and a log Bayes’ factor of more than 3 is considered to be a strong support in favor of the alternative network model ($H_2$) [265].
Chapter 7

Conclusions

Traditional epidemiological theory suggests that socially complex species that live in large groups have elevated risks of disease transmission due to higher frequency of contacts between hosts [39, 199, 200]. Group living species, therefore, resort to behavioral, physiological defense mechanisms to offset their disease transmission risks [29, 43, 223]. However, a growing body of theoretical literature on network epidemiology suggests that the organization of social contacts plays a central role in predicting disease outcomes in host populations [40, 41, 42, 46]. In this dissertation, I investigate how and under what conditions social structure of animal populations reduces the spread of pathogens. Contrary to expectation, my study does not support the hypothesis that subgroup structure in animal populations modulates the association between group size and infectious disease risk, unless the populations are highly subdivided. Highly subdivided populations are able to reduce the global spread of infectious disease by structurally trapping infectious disease within local modules. I also provides evidence that only a few unique features of social structure are predictive of disease outcomes in different social systems.

In this dissertation, I advance modeling techniques of wildlife infectious diseases by developing novel tools and methods that incorporate traditional data sampling schemes into network modeling. In particular, the results of this study address three common challenges currently faced in the field of wildlife epidemiology. The first challenge is quantifying the social structure of species where direct observations of
interactions are either rare or difficult. I show that for such species, spatially-driven bipartite network approaches can help infer covert social associations between individuals in a population. The network modeling technique applied in the study can also be used by conservation biologists and wildlife managers to mitigate the risks of sudden die-offs due to natural or anthropogenic threats on wild animal populations. In addition, this approach can be important to evaluate epidemiological implications of management strategies, such as translocation, that are known to disturb social interactions in animal groups. Highly visited refuges can be used as early warning sentinels to monitor population health, and may serve as an efficient alternative to traditional demographic metrics (such as mortality rate, reproduction rate).

The second challenge addressed in this dissertation is the fact that network models are data intensive and collecting data at an individual scale is often time-consuming or expensive [266]. Continuous observations of individuals in wildlife populations are often impractical, and thus the datasets collected typically contain missing observations. In such cases, I show that the semi-empirical "null" networks can provide important insights into the impact of key local and global structures on the dynamics of infection spread through host populations. Additionally, null network models can be used to serve as random controls when investigating the functional role of complex network features beyond the byproduct of group size and average connectivity in empirical networks. Null networks also have the ability to isolate the network feature of interest from other correlated structures found in empirical systems. Analysis of null networks therefore provides key insights into the evolutionary outcomes and functional role of unique local and global structural features observed in biological, ecological, and social networks.

The third challenge that I address is that transmission pathways are often unknown for wildlife infectious diseases. This precludes an exact definition of disease-
causing contact which is central to constructing accurate contact network models. To address this gap I develop a tool, called INoDS, that estimates the predictive power of empirical contact networks in explaining the spread of infectious diseases, and enables hypothesis testing between different contact network models. This tool therefore can be used to (i) identify transmission mode of pathogen spread in wildlife and animal populations without relying on laboratory experimentation, (ii) establish the predictive power of proxy measures of contact in cases where limited interaction data are available, and (iii) establish an association between contact intensity and the risk of infection transfer.

Overall, this study provides much needed insight into the long-standing and fundamental questions about the associations between social systems, contact network structure and infectious disease spread in wildlife populations. The tools and data analysis methods I have developed in this work can be easily extended to study human contact networks and other biological and non-biological complex networks. The results of this dissertation, therefore, also provides significant resources for the broader network research community.
## Appendix A

# Chapter 2 Supplementary materials

Table A.1: Location, surveyed years, total tortoises tracked and average observations at the study sites.

<table>
<thead>
<tr>
<th>Site</th>
<th>Site abbr</th>
<th>County, State</th>
<th>Survey years</th>
<th>Total tortoises tracked</th>
<th>Average observations (± standard deviation) per month per tortoise</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bird Spring</td>
<td>BSV</td>
<td>Clark, NV</td>
<td>1996-99</td>
<td>237</td>
<td>Active: 3.94 ± 1.72, Inactive: 1.82 ± 1.05</td>
</tr>
<tr>
<td>Coyote Springs</td>
<td>CS</td>
<td>Clark, NV</td>
<td>2006-14</td>
<td>71</td>
<td>Active: 3.92 ± 3.54, Inactive: 1.51 ± 0.71</td>
</tr>
<tr>
<td>Fort Irwin</td>
<td>FI</td>
<td>Barstow, CA</td>
<td>2005-14</td>
<td>2138</td>
<td>Active: 3.43 ± 3.03</td>
</tr>
<tr>
<td>Halfway</td>
<td>HW</td>
<td>Lincoln, NV</td>
<td>2008-12</td>
<td>49</td>
<td>Inactive: 1.66 ± 1.27, Active: 2.87 ± 3.30</td>
</tr>
<tr>
<td>Lake Meade</td>
<td>LM</td>
<td>Clark, NV</td>
<td>1999-00</td>
<td>53</td>
<td>Inactive: 1.43 ± 0.59, Active: 3.11 ± 1.27</td>
</tr>
<tr>
<td>McCullough</td>
<td>MC</td>
<td>Clark, NV</td>
<td>2012-13</td>
<td>27</td>
<td>Active: 3.89 ± 1.21</td>
</tr>
<tr>
<td>Pass</td>
<td>PV</td>
<td>Clark, NV</td>
<td>2001-13</td>
<td>92</td>
<td>Inactive: 1.29 ± 0.48, Active: 3.99 ± 2.37</td>
</tr>
<tr>
<td>Pinte Valley</td>
<td>PV</td>
<td>Clark, NV</td>
<td>13</td>
<td>13</td>
<td>Active: 4.35 ± 1.87</td>
</tr>
<tr>
<td>St. George</td>
<td>SG</td>
<td>Washington, UT</td>
<td>1998-00</td>
<td>62</td>
<td>Inactive: 1.53 ± 0.82, Active: 4.35 ± 1.87</td>
</tr>
<tr>
<td>Stateline Pass</td>
<td>SL</td>
<td>Clark, NV</td>
<td>2012-13</td>
<td>15</td>
<td>Inactive: 1.71 ± 1.17, Active: 3.36 ± 1.02</td>
</tr>
</tbody>
</table>

Summary of the nine sites across the desert tortoise habitat used in this study
Figure A.1: Standardized Pearson residuals of the burrow switching model.
Figure A.2: Standardized Pearson residuals of the burrow popularity model.
Figure A.3: Effect of a) sex/age class, b) sampling period × seasonal rainfall (rain, measured in inches), c) local tortoise density and d) local burrow density on burrow switching levels of desert tortoises. Error bars represent 95% confidence intervals, and different letters above the points denote a significant difference between the means (P < 0.05).
Figure A.4: Popularity level of burrows in various a) surface roughness, b) burrow age, c) sampling periods, and d) density conditions. Note that the scale of the y-axis in panel (d) is different. An average of 3 and 6.5 burrows around focal burrows is denoted as low and high burrow density respectively. Error bars represent 95% confidence intervals, and different letters above the points denote a significant difference between the means (P < 0.05).
Table A.2: Location of the weather stations used to obtain climatic information for the nine study sites.

<table>
<thead>
<tr>
<th>Site</th>
<th>County, state</th>
<th>Weather Station</th>
<th>Latitude</th>
<th>Longitude</th>
</tr>
</thead>
<tbody>
<tr>
<td>BSV</td>
<td>Clark, NV</td>
<td>Overton, NV</td>
<td>35.551°N</td>
<td>114.458°W</td>
</tr>
<tr>
<td>CS</td>
<td>Clark, NV</td>
<td>Overton, NV</td>
<td>35.551°N</td>
<td>114.458°W</td>
</tr>
<tr>
<td>FI</td>
<td>Barstow, CA</td>
<td>Barstow Dagget Airport, CA</td>
<td>34.854°N</td>
<td>116.786°W</td>
</tr>
<tr>
<td>HW</td>
<td>Lincoln, NV</td>
<td>Pahrangat, NV</td>
<td>37.269°N</td>
<td>115.122°W</td>
</tr>
<tr>
<td>LM</td>
<td>Clark, NV</td>
<td>Overton, NV</td>
<td>35.551°N</td>
<td>114.458°W</td>
</tr>
<tr>
<td>MC</td>
<td>Clark, NV</td>
<td>Overton, NV</td>
<td>35.551°N</td>
<td>114.458°W</td>
</tr>
<tr>
<td>PV</td>
<td>Clark, NV</td>
<td>Overton, NV</td>
<td>35.551°N</td>
<td>114.458°W</td>
</tr>
<tr>
<td>SG</td>
<td>Washington, UT</td>
<td>St George, UT</td>
<td>37.107°N</td>
<td>113.561°W</td>
</tr>
<tr>
<td>SL</td>
<td>Clark, NV</td>
<td>Overton, NV</td>
<td>35.551°N</td>
<td>114.458°W</td>
</tr>
</tbody>
</table>

Figure A.5: Effect of population stressors of a) translocation, b) URTD, and c) winter rainfall on burrow switching levels of desert tortoises. In panel a), C = Controls, R = Residents, T = Translocated, ER = Ex-Residents and ET= Ex-Translocated. Unhealthy animals in panel b) represents individuals exhibiting clinical signs of the infection and healthy represents uninfected or asymptomatic tortoises. Error bars represent 95% confidence intervals, and different letters above the points denote a significant difference between the means (P < 0.05).
Table A.3: Network density, degree homophily, clustering coefficient, modularity and degree centralization of tortoise social networks at five study sites with control animals. Tortoise social network is constructed as the single mode projections of bipartite network of burrow use for active (Mar - Oct) and inactive (Nov-Feb) season of each surveyed year. For consistent comparison, networks with less than five edges were excluded from the analysis.

<table>
<thead>
<tr>
<th>Site</th>
<th>Season</th>
<th>Total networks</th>
<th>Network density</th>
<th>Degree homophily</th>
<th>Clustering coefficient</th>
<th>Modularity</th>
<th>Degree centralization</th>
</tr>
</thead>
<tbody>
<tr>
<td>CS</td>
<td>Active</td>
<td>8</td>
<td>0.151 (8, 0.05)</td>
<td>0.376 (3, 0.05)</td>
<td>0.382 (7, 0.12)</td>
<td>0.573 (7, 0.13)</td>
<td>0.169 (8, 0.05)</td>
</tr>
<tr>
<td></td>
<td>Inactive</td>
<td>2</td>
<td>0.172 (2, 0.01)</td>
<td>1.0 (2, 0.00)</td>
<td>0.467 (2, 0.19)</td>
<td>0.677 (2, 0.02)</td>
<td>0.072 (2, 0.03)</td>
</tr>
<tr>
<td></td>
<td>Active</td>
<td>3</td>
<td>0.193 (3, 0.05)</td>
<td>0.588 (1, -)</td>
<td>0.230 (3, 0.21)</td>
<td>0.642 (1, -)</td>
<td>0.149 (3, 0.06)</td>
</tr>
<tr>
<td>HW</td>
<td>Inactive</td>
<td>0</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>Active</td>
<td>2</td>
<td>0.266 (2, 0.01)</td>
<td>0.243 (2, 0.49)</td>
<td>0.592 (2, 0.25)</td>
<td>0.338 (1, -)</td>
<td>0.264 (1, -)</td>
</tr>
<tr>
<td>MC</td>
<td>Inactive</td>
<td>0</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>Active</td>
<td>7</td>
<td>0.184 (8, 0.06)</td>
<td>0.484 (4, 0.36)</td>
<td>0.350 (7, 0.23)</td>
<td>0.601 (4, 0.07)</td>
<td>0.158 (7, 0.06)</td>
</tr>
<tr>
<td>PV</td>
<td>Inactive</td>
<td>0</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>Active</td>
<td>2</td>
<td>0.285 (2, 0.10)</td>
<td>0 (1, -)</td>
<td>0.513 (2, 0.17)</td>
<td>0.416 (2, 0.14)</td>
<td>0.250 (2, 0.10)</td>
</tr>
<tr>
<td>SL</td>
<td>Inactive</td>
<td>0</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

Values in parentheses are sample size of individual metric and standard deviation. Only statistically significant values with P < 0.05 were used to calculate averages and standard deviation.
Table A.4: Correlation between geographical distances and edge occurrence in tortoise social networks during active season of each surveyed year. Tortoise social networks with less than five edges were excluded from the analysis.

<table>
<thead>
<tr>
<th>Site</th>
<th>Year</th>
<th>Nodes</th>
<th>Edges</th>
<th>Edge correlation</th>
<th>P value</th>
</tr>
</thead>
<tbody>
<tr>
<td>CS</td>
<td>2006</td>
<td>9</td>
<td>7</td>
<td>-0.439</td>
<td>0.034</td>
</tr>
<tr>
<td>CS</td>
<td>2007</td>
<td>21</td>
<td>35</td>
<td>-0.318</td>
<td>0.001</td>
</tr>
<tr>
<td>CS</td>
<td>2008</td>
<td>17</td>
<td>35</td>
<td>-0.384</td>
<td>0.005</td>
</tr>
<tr>
<td>CS</td>
<td>2009</td>
<td>22</td>
<td>27</td>
<td>-0.350</td>
<td>0.000</td>
</tr>
<tr>
<td>CS</td>
<td>2010</td>
<td>24</td>
<td>35</td>
<td>-0.400</td>
<td>0.000</td>
</tr>
<tr>
<td>CS</td>
<td>2011</td>
<td>18</td>
<td>14</td>
<td>-0.322</td>
<td>0.000</td>
</tr>
<tr>
<td>CS</td>
<td>2012</td>
<td>17</td>
<td>20</td>
<td>-0.330</td>
<td>0.002</td>
</tr>
<tr>
<td>CS</td>
<td>2013</td>
<td>18</td>
<td>16</td>
<td>-0.222</td>
<td>0.033</td>
</tr>
<tr>
<td>HW</td>
<td>2009</td>
<td>11</td>
<td>9</td>
<td>-0.581</td>
<td>0.000</td>
</tr>
<tr>
<td>HW</td>
<td>2011</td>
<td>9</td>
<td>6</td>
<td>-0.658</td>
<td>0.000</td>
</tr>
<tr>
<td>HW</td>
<td>2012</td>
<td>8</td>
<td>7</td>
<td>-0.895</td>
<td>0.000</td>
</tr>
<tr>
<td>MC</td>
<td>2012</td>
<td>14</td>
<td>25</td>
<td>-0.586</td>
<td>0.000</td>
</tr>
<tr>
<td>MC</td>
<td>2013</td>
<td>13</td>
<td>20</td>
<td>-0.490</td>
<td>0.003</td>
</tr>
<tr>
<td>PV</td>
<td>2005</td>
<td>10</td>
<td>13</td>
<td>-0.653</td>
<td>0.001</td>
</tr>
<tr>
<td>PV</td>
<td>2008</td>
<td>14</td>
<td>11</td>
<td>-0.392</td>
<td>0.010</td>
</tr>
<tr>
<td>PV</td>
<td>2009</td>
<td>8</td>
<td>5</td>
<td>-0.513</td>
<td>0.005</td>
</tr>
<tr>
<td>PV</td>
<td>2010</td>
<td>12</td>
<td>21</td>
<td>-0.563</td>
<td>0.009</td>
</tr>
<tr>
<td>PV</td>
<td>2011</td>
<td>15</td>
<td>12</td>
<td>-0.402</td>
<td>0.000</td>
</tr>
<tr>
<td>PV</td>
<td>2012</td>
<td>20</td>
<td>26</td>
<td>-0.469</td>
<td>0.000</td>
</tr>
<tr>
<td>PV</td>
<td>2013</td>
<td>16</td>
<td>16</td>
<td>-0.494</td>
<td>0.000</td>
</tr>
<tr>
<td>SL</td>
<td>2012</td>
<td>6</td>
<td>5</td>
<td>-0.777</td>
<td>0.007</td>
</tr>
<tr>
<td>SL</td>
<td>2013</td>
<td>11</td>
<td>13</td>
<td>-0.484</td>
<td>0.004</td>
</tr>
</tbody>
</table>
Table A.5: Potential interactions considered for the burrow switching and the burrow popularity model.

<table>
<thead>
<tr>
<th>Model</th>
<th>dLogLik</th>
<th>dBIC</th>
<th>df</th>
<th>weight</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Burrow switching model</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fixed effects + sampling period×sex/age class, sampling period×seasonal rainfall, local tortoise density×local burrow density</td>
<td>58.3</td>
<td>76.5</td>
<td>39</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Fixed effects + sampling period×sex/age class, sampling period×seasonal rainfall</td>
<td>58.2</td>
<td>67.4</td>
<td>38</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Fixed effects + sampling period×sex/age class, local tortoise density×local burrow density</td>
<td>12.1</td>
<td>122.5</td>
<td>34</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Fixed effects + sampling period×seasonal rainfall, local tortoise density×local burrow density</td>
<td>45.6</td>
<td>9.1</td>
<td>29</td>
<td>0.011</td>
</tr>
<tr>
<td>Fixed effects + sampling period×sex/age class</td>
<td>11.8</td>
<td>113.9</td>
<td>33</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Fixed effects + sampling period×seasonal rainfall</td>
<td>45.5</td>
<td>0.0</td>
<td>28</td>
<td>0.989</td>
</tr>
<tr>
<td>Fixed effects + local tortoise density×local burrow density</td>
<td>0.3</td>
<td>53.0</td>
<td>24</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Fixed effects only</td>
<td>0.0</td>
<td>44.4</td>
<td>23</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td><strong>Burrow popularity model</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fixed effects + sampling period×local tortoise density, sampling period×seasonal rainfall, local tortoise density×local burrow density</td>
<td>186.9</td>
<td>4.8</td>
<td>41</td>
<td>0.082</td>
</tr>
<tr>
<td>Fixed effects + sampling period×local tortoise density, sampling period×seasonal rainfall</td>
<td>113.6</td>
<td>141.5</td>
<td>40</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Fixed effects + sampling period×local tortoise density, local tortoise density×local burrow density</td>
<td>164.2</td>
<td>0.0</td>
<td>36</td>
<td>0.918</td>
</tr>
<tr>
<td>Fixed effects + sampling period×seasonal rainfall, local tortoise density×local burrow density</td>
<td>102.5</td>
<td>123.4</td>
<td>36</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Fixed effects + sampling period×local tortoise density</td>
<td>91.4</td>
<td>135.6</td>
<td>35</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Fixed effects + sampling period×seasonal rainfall</td>
<td>17.3</td>
<td>283.8</td>
<td>35</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Fixed effects + local tortoise density×local burrow density</td>
<td>83.3</td>
<td>111.6</td>
<td>31</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Fixed effects only</td>
<td>0.0</td>
<td>268.1</td>
<td>30</td>
<td>&lt;0.001</td>
</tr>
</tbody>
</table>

Fixed effects include all the predictor variables described methods.
Appendix B

Chapter 3 Supplementary materials

B.1 Expected modularity

The strength of community structure in a network with $K$ partitions is defined as

$$Q = \sum_{k=1}^{K} (e_{kk} - a_k^2)$$  \hspace{1cm} (B.1)

where $e_{kk}$ denotes the fraction of edges within the module $k$ and $a_k$ is the fraction of the total edges of nodes of module $k$.

Now, if $\overline{d_k}$ is the average-degree of the module $k$, $\overline{d^w_k}$ is its average within-module degree and $s_k$ is the total number of nodes in the module, then equation S1 can be written as:

$$Q = \sum_{k=1}^{K} \left[ \frac{\overline{d^w_k}s_k}{\bar{d}n} - \left( \frac{\overline{d^k}s_k}{\bar{d}n} \right)^2 \right]$$  \hspace{1cm} (B.2)

where $\overline{d}$ is the average degree of the network and the network size and total modules is $n$ and $K$ respectively. If the average-degree of each module is equal to the average degree of the network, i.e. $\overline{d_1} = \overline{d^2} = \ldots = \overline{d^K} = \overline{d}$, and if the average within-degree of each module is equal to the average within-degree of the network, i.e. $\overline{d^w_1} = \overline{d^w_2} = \ldots = \overline{d^w_K} = \overline{d^w}$, then equation (S2) can be written as:

$$Q = \sum_{k=1}^{K} \left[ \frac{\overline{d^w}s_k}{\bar{d}n} - \left( \frac{s_k}{n} \right)^2 \right] = \frac{\overline{d^w}}{\bar{d}} - \sum_{k=1}^{K} \left( \frac{s_k}{n} \right)^2$$  \hspace{1cm} (B.3)
where $\overline{d_w}$ is the average within-module degree of the network. Now, if all the modules are of equal sizes, equation (S3) can be further reduced to:

$$Q = \frac{\overline{d_w}}{\overline{d}} - \frac{1}{K}$$  \hspace{1cm} (B.4)

Thus, the expected modularity in this case can be expressed in terms of the ratio of average module degree $\overline{d_w}$ and average total degree $\overline{d}$ of the network, as well as the total number of partitions or modules $K$ in the network.

### B.2 Tolerance on average-degree and average within-degree of individual modules

We note that Equation S3 can be used to estimate modularity only when module-level average degree and average within-degree match up to the overall network average-degree and average within-degree, i.e., $\overline{d^1} = \overline{d^2} = ... = \overline{d^K} = \overline{d}$ and $\overline{d^1_w} = \overline{d^2_w} = ... = \overline{d^K_w} = \overline{d_w}$. To ensure that these conditions are valid we used rejection sampling of both degree and within-degree sequence. We define the tolerance on the expected modularity, $\epsilon$, to be 0.01 and calculate the tolerance on within-degree and degree sequence as follows:

Let $\epsilon_{d_w}$ be the tolerance on sampled within-degree sequence. We define $\epsilon_d = 0.5\epsilon_{d_w}$ to be the tolerance on degree sequence. From Eq. (S2), the observed modularity can be thus written as:

$$\hat{Q} = \sum_{k=1}^{K} \left[ \frac{\overline{d_w} \pm \epsilon_{d_w}}{\overline{d}.n} \cdot s_k - \left( \frac{\overline{d} \pm \epsilon_d}{\overline{d}.n} \right) \cdot s_k \right] = \sum_{k=1}^{K} \left[ \frac{\overline{d_w}.s_k \pm \epsilon_{d_w}.s_k}{\overline{d}.n} - \frac{\overline{d}.s_k^2}{\overline{d}.n^2} \pm 2.\frac{\epsilon_d.\overline{d}.s_k^2}{\overline{d}.n^2} \right]$$  \hspace{1cm} (B.5)

By ignoring the $\epsilon_d^2$ term which is negligible. $\hat{Q}$ can be further simplified to:
Figure B.1: Degree distribution of the four biological networks. The within-degree distribution roughly follows the total degree distribution in all of the networks.

\[ \hat{Q} = Q \pm \sum_{k=1}^{K} \left[ \frac{\epsilon_{dw} \cdot s_k}{d \cdot n} - \frac{2 \cdot \epsilon_d \cdot d \cdot s^2}{d^2 \cdot n^2} \right] = Q \pm \frac{\epsilon_{dw}}{d} - \sum_{k=1}^{K} \left[ \frac{\epsilon_{dw} \cdot s^2_k}{d \cdot n^2} \right] \quad (B.6) \]

as \( \hat{Q} - Q = \epsilon = 0.01 \), \( \epsilon_{dw} \) can be thus calculated as

\[ 0.01 = \frac{\epsilon_{dw}}{d} \left[ 1 - \sum_{k=1}^{K} \frac{s^2_k}{n^2} \right] \quad (B.7) \]

B.3 Within-module degree distribution follows the total degree distribution

In Figure B.1 we plotted the probability density of the total-degree and within-degree distribution for four empirical biological networks namely a) Metabolic interaction
network of *Caenorhabditis elegans* [163]; b) Food web, depicting the network of trophic interactions at Little Rock Lake in Wisconsin [161]; c) Protein interaction in Yeast [162]; and d) Network of social interactions in a community of 62 dolphins living off Doubtful Sound, New Zealand [164]. We found that the within-degree distribution of most of the empirical networks closely follows the network’s total degree distribution indicating a fractal like behavior of the network. Based on this observation we limited our discussions to modular random networks which have similar within-module and total degree distribution. However, our model can be extended to allow for arbitrary within-degree distributions or sequences. To demonstrate, we generated examples of graphs with arbitrary within-degree distributions (Table S1; fourth, fifth and sixth network type) and compared their network properties to modular graphs with similar degree and within-degree distributions (Table S1; first, second and third network type). The modularity value of all generate random graphs was fixed at 0.2. We found that the network properties of clustering coefficient and average path length to be similar across all the network types (Table S1). Degree assortativity value is close to zero for all network types except for graphs with Poisson degree distribution and geometric within-degree distribution where edge connections are constrained.

**B.4 Rejection rate of degree and within-degree sequence**

Here we estimate the rejection rate of sampling degree and within-degree sequences during the generation of 2000 nodes networks with mean degree 10. The rejection rates are calculated based on the number of times each sequence is rejected per graph generation process. Average rejection rate is calculated over 50 such generation process. Figure B.2 shows the expectation of the rejection rate, which we estimate by sampling the average rejection rate ten times. As expected, the rejection rate of
Figure B.2: Rejection rate of degree and within-degree sequence sampling for random modular graphs with geometric degree distribution and (a-c) Poisson degree distribution (d-f) Geometric degree distribution. Rejection rate is estimated over three modularity values. The generated graphs have 2000 nodes, mean network degree of 10, and consist of 10 communities.

The rejection rate of within-degree sequence increases with network modularity.

B.5 Generating disassortative modular random graphs

Anti-modular or disassortative modular random graphs are graphs in which nodes tend to connect to nodes of other modules. This results in within-module edge density to be less than what is expected at random and the value of modularity coefficient, $Q$, to be negative. In Figure B.3 we generate both anti-modular and modular random graphs with identical size ($n = 150$), average degree ($\bar{d} = 5$), number of modules ($K = 3$) and degree distribution (power-law). The absolute $Q$ value of both the...
Figure B.3: Generating disassortative modular random graphs: Modular random graphs with \( n = 150 \), \( m = 375 \), \( K=3 \), \( P(s=50)=1 \) and \( p_k \) is power law with modularity values of: a) \( Q= -0.2 \) and b) \( Q= 0.2 \). In anti-modular (disassortative) graphs the between-module edge density is more than within-module edge density, whereas the opposite is true in modular random graphs.
graphs is identical (i.e. $|Q| = 0.2$), but anti-modular graph (Fig S3a) has between-module edge density higher than within-module edge density, whereas the opposite is true in the modular random graph (Fig S3b).

**B.6 Comparing structural properties of modular random and SBM graphs**

Here we compare the structural properties of modular random graphs generated by our model to the ones generated by degree-corrected stochastic block models (DC-SBM) as described in [133]. SBM is defined by a $k \times k$ stochastic block matrix, where $k$ is the number of modules and $M_{ij}$ gives the probability that a node of module $i$ is connected to a node of module $j$. The DC-SBM version further defines a propensity parameter $\gamma_u$ that controls the expected degree of node $u$. 

---

**Figure B.4:** Comparing modular random graphs and stochastic block model (SBM) graphs: Network property of (a) degree assortativity, (b) clustering coefficient, and (c) path length in modular random graphs (MRG) and stochastic block model (SBM) graphs of 2000 nodes, mean degree 10 and 10 modules on increasing modularity ($Q$). Each data point represents the average value of 50 random graphs. The module size of each graph follows a Poisson distribution with mean size of 200. Data points for path length of SBM geometric networks is missing as the generated networks are disconnected. Standard deviations are plotted as error bars.
We used a Python module (graph-tool) to generated SBM graphs. Since a formal relationship between the SBM parameters and modularity does not exist, we manually adjusted the parameters values to achieve the desired level of modularity and network parameters. Figure B.4 shows two types of SBM graphs: (a) random graphs with Poisson degree-distribution and Poisson within-degree distribution, and (b) random graphs with geometric degree and within-module degree-distribution. Using the Python module and desired network parameters, we were able to generate graphs with a maximum modularity value of 0.4 for both these network types. We therefore generated fifty random graphs at each level modularity and estimated the average values of degree assortativity (Figure B.4a), clustering coefficient (Figure B.4b) and path length (Figure B.4c) of these graphs. The module size follows a Poisson distribution in each of these graphs. To compare DC-SBM to the modular random graphs generated by our model, we generated graphs with identical network parameters and modularity values and report their network properties was well. As Figure B.4 shows the structural properties of the graphs generated from the graph-tool Python module are similar to those generated by our algorithm. We note, however, that this is a limited comparison and highly dependent on the implementation of the SBM in graph-tool. As discussed in the Previous Work section of the main article, full use of the SBM for generating benchmark or null networks remains to be fully explored.

B.7 Effect of network size on network properties of modular random graphs

Here we varied the network size keeping the ratio of community size to the total network size (i.e \( s/n = 0.1 \)) constant. As each network comprised 10 communities, increase in total network size also corresponds to the increase in average community.
Figure B.5: Network property of (a) degree assortativity, (b) clustering coefficient and (c) path length in random modular graphs of mean degree 5 and 10 modules on increasing network size ($n$). Each data point represents the average value of 50 random graphs. Standard deviations are plotted as error bars.

We observed that, except for very small networks, the assortativity coefficient remains close to zero for all network size (Figure B.5a). The negative degree correlation for small networks can be explained by the structural degree cut-off constraint in the communities, i.e. indegree of nodes in a community can attain a value of at-most equal to its community sizes ($\max(w_d) \leq n_c$). For smaller networks, the highest value of $w_d$ is constrained by the small average community size, which results in the total number of high indegree to be much less than expected. Thus, during the randomization step the high indegree nodes connect much more to the low degree nodes which result in disassortative network. A similar observation was noted in hierarchically modular networks by Jing [145]. Clustering coefficient is higher for small networks but decreases to a value close to zero in networks with more than 400 nodes, which is observed in larger networks as well (Figure B.5b). As expected, the average shortest path length increases proportionally with network size (Figure B.5c).
Figure B.6: Network property of (a) degree assortativity, (b) clustering coefficient and (c) path length in random modular graphs with 10 modules over a range of mean network degree. Each network has 1000 nodes. The data point represents the average value of 50 random graphs. Standard deviations are plotted as error bars.

B.8 Effect of average network degree on network properties of modular random graphs

We next tested the effect of network mean degree on other properties of the network. We observed that geometric and power-law null modular networks become disassortative with higher $d$ value, while Poisson networks do not show any assortative interaction at any value $\bar{d}$ (Figure B.6a). The tendency of geometric and power-law null modular networks to become disassortative could again be due to the structural cut-off constraint of nodal indegrees. As the average degree increases, the graph becomes more dense and hence creates more implicit triangles, resulting in a gradual increase in clustering (Figure B.6b). Decrease in average shortest path length with increase in mean network degree is also well known [146, 147].
Figure B.7: Network property of (a) degree assortativity, (b) clustering coefficient and (c) path length in random modular graphs of size 1000 with mean degree 10 but different number of modules. As the total network size is fixed (=1000) and each module in a network is of equal size, increasing the number of modules in a network corresponds to a decrease in average community size. Each data point represents the average value of 50 random graphs. Standard deviations are plotted as error bars.

B.9 Effect of average community size on network properties of modular random graphs

We also investigated the effect of average community size on the network properties of the null modular network. Figure B.7 summarizes the results for networks with a network size of 1000 but different number of modules. A smaller number of modules thus corresponds to a larger average community size. We observed that the community size does not effect the assortative interaction for Poisson networks (Figure B.7a). Geometric and power-law networks show disassortative interactions in networks with small community size due to structural degree cut-off constraint explained above. The density of edges within smaller communities is high, which causes high clustering (Figure B.7b). However, the average shortest path length is unaffected by the community size (Figure B.7c) as the total network size and network mean degree is constant across all network types.
Figure B.8: Performance of various community detection algorithms on random modular networks with Poisson degree distribution. Network size $n = 2000$, mean degree $\langle d \rangle = 10$, number of modules $m = 10$. Each data point represents the average results of 25 detection runs on a generated modular random network. For each $Q$ value 10 modular random networks were generated.

B.10 Performance of other community detection algorithms on modular random graphs

Here we estimated the modularity of our generated random modular Poisson (Figure B.8), geometric (Figure B.9), and power-law (Figure B.10) networks using four additional community detection algorithms namely: (a) spinglass or Potts model [151]; (b) walktrap algorithm [154], (c) Infomap algorithm [152], and (d) label propagation model [153]. Overall, the accuracy of these algorithms improves with increasing $Q$ value.
Figure B.9: Performance of various community detection algorithms on random modular networks with geometric degree distribution. Network size $n=2000$, mean degree $(\bar{d})=10$, number of modules $(m)=10$. Each data point represents the average results of 25 detection runs on a generated modular random network. For each $Q$ value 10 modular random networks were generated.
Figure B.10: Performance of various community detection algorithms on random modular networks with power-law degree distribution. Network size $n = 2000$, mean degree $\langle d \rangle = 10$, number of modules $m = 10$. Each data point represents the average results of 25 detection runs on a generated modular random network. For each $Q$ value 10 modular random networks were generated.
We tested the accuracy of network partitioning by Louvain and fast modularity algorithm (Figure B.11) in random modular networks with a mean network degree of 10 using Jaccard similarity index ($J$) and variation of information (VI) as a measure of similarity. Jaccard index is the ratio of the number of nodes classified in the same module by both the partitions to the total number of nodal pairs, i.e.
\[ J = \frac{w_{11}}{w_{11} + w_{00} + w_{10}} \]  

(B.8)

where \( w_{11} \) represent the number of nodal pairs that are in the same module for both the partitions, \( w_{00} \) are the nodal pairs that are in different modules in both the partitions and \( w_{10}(w_{01}) \) are the number of pairs that are put together in the same module by one partition but not by the other. The value of Jaccard index ranges from 0 to 1, with 1 indicating a perfect partition match.

VI measures the amount of information lost and gained in changing from clustering \( C \) to clustering \( C' \) [156] and is defined as

\[
VI(C, C') = H(C|C') + H(C'|C) \tag{B.9}
\]

or,

\[
VI(C, C') = [H(C) - I(C|C') - [H(C') - I(C|C')] \tag{B.10}
\]

where \( H(C) \) and \( H(C') \) represents uncertainty in cluster \( C \) and \( C' \) respectively, and \( I(C|C') \) is the mutual information between the two clustering. In other words, the first term of equation (S10) measures the amount of information that we loose, while the second term measures the amount of information that we gain, when going to clustering \( C' \) from \( C \).

B.12 Null analysis of empirical networks

We generated random modular graphs for each of the four biological networks by randomizing the within-edge and between-edge connections. Specifically, we generated 50 such random graphs using the estimates of total degree distribution, within-degree distribution, and distribution of module size \( P(\bar{s}) \) as the empirical network but used our model to connect the within- and between-edges. We next measured networks
properties such as clustering \((C)\), average path length \((L)\), assortativity \((r)\) for each of the random network and computed the ensemble mean. Table 1 records the value of each of these properties for the empirical networks and the relative deviation of the ensemble mean of random modular graphs from the observed value (i.e. deviation \(= \frac{\text{observed value} - \text{ensemble mean}}{\text{observed value}}\))

Table B.1: Comparisons of random modular networks with different degree- and within-degree distributions type.

<table>
<thead>
<tr>
<th>Network types</th>
<th>Modularity, (Q)</th>
<th>Degree assortativity, (r)</th>
<th>Clustering coefficient, (C)</th>
<th>Path length (L)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1) Poisson degree distribution, Poisson within-degree distribution</td>
<td>0.2 ([0.004])</td>
<td>0.02 ([0.009])</td>
<td>0.01 ([0])</td>
<td>3.56 ([0.002])</td>
</tr>
<tr>
<td>2) Geometric degree distribution, Geometric within-degree distribution</td>
<td>0.2 ([0.003])</td>
<td>-0.03 ([0.009])</td>
<td>0.02 ([0.001])</td>
<td>3.48 ([0.010])</td>
</tr>
<tr>
<td>3) Power-law degree distribution, Power-law within-degree distribution</td>
<td>0.2 ([0.002])</td>
<td>-0.01 ([0.009])</td>
<td>0.01 ([0])</td>
<td>3.48 ([0.009])</td>
</tr>
<tr>
<td>4) Poisson degree distribution, Geometric within-degree distribution</td>
<td>0.2 ([0.003])</td>
<td>0.13 ([0.009])</td>
<td>0.007 ([0])</td>
<td>3.57 ([0.002])</td>
</tr>
<tr>
<td>5) Geometric degree distribution, Power-law within-degree distribution</td>
<td>0.2 ([0.002])</td>
<td>-0.03 ([0.009])</td>
<td>0.02 ([0.001])</td>
<td>3.49 ([0.013])</td>
</tr>
<tr>
<td>6) Geometric degree distribution, Poisson within-degree distribution</td>
<td>0.2 ([0.004])</td>
<td>0.03 ([0.009])</td>
<td>0.01 ([0.001])</td>
<td>3.50 ([0.011])</td>
</tr>
</tbody>
</table>

Network property of assortativity, clustering coefficient, and path length in random modular graphs of size 2000 with mean degree 10. Each network type represents random modular graphs with a specific degree and within-degree distribution. Module sizes of all the generated networks follow a Poisson distribution. Each value represents an average of 50 random graphs. Standard deviations are included within square brackets.
Table B.2: Comparisons of empirical and random networks with randomized within-edge and between-edge connections.

<table>
<thead>
<tr>
<th>Biological Network Type</th>
<th>N</th>
<th>k</th>
<th>Q</th>
<th>C</th>
<th>L</th>
<th>r</th>
</tr>
</thead>
<tbody>
<tr>
<td>Little Rock Foodweb Interactions</td>
<td>183</td>
<td>26.79</td>
<td>0.36</td>
<td>0.32 [-88%]</td>
<td>2.15 [1%]</td>
<td>-0.26 [69%]</td>
</tr>
<tr>
<td>Yeast Protein Interactions</td>
<td>4713</td>
<td>6.31</td>
<td>0.54</td>
<td>0.09 [-44%]</td>
<td>-</td>
<td>-0.14 [228%]</td>
</tr>
<tr>
<td>C.elegans Metabolic Interactions</td>
<td>453</td>
<td>9.01</td>
<td>0.44</td>
<td>0.65 [22%]</td>
<td>2.66 [1%]</td>
<td>-0.22 [0]</td>
</tr>
<tr>
<td>Dolphin Social Interaction</td>
<td>62</td>
<td>5.13</td>
<td>0.52</td>
<td>0.26 [-12%]</td>
<td>3.36 [1%]</td>
<td>-0.04 [450%]</td>
</tr>
</tbody>
</table>

For each of the four empirical network we generated 50 null modular network constrained to have the same total-, within- and between-degree list as the empirical network. The table summarizes network statistics of empirical network viz. the network size (N), average network degree (k), modularity(Q), clustering (C), average shortest path-length (L) and degree assortativity (r). The value in brackets is the relative deviation of ensemble mean of null modular networks from the observed value. The path length value for the empirical Yeast-Protein interaction network is missing as the network is not fully-connected.
C.1 Animal social networks

We obtained previously published (and publicly available) social networks of 69 groups across 43 animal species. The networks were selected based on the criterion that edge connections can serve as actual routes of infectious disease spread. For groups that had multiple temporal snapshots of social network, we randomly selected one network for the analysis to avoid intra-group correlation in network metrics. Few networks had weighted edges. To enable consistent comparison across all networks, we transformed the edges of these networks to binary edges by employing an edge filtering approach [203]. To this end, edges with weights less than 50th percentile of edge weights were removed from the network, and the remaining edges were assigned an edge weight of one.

C.2 Measure of relative modularity

We used modularity ($Q$) proposed by Newman [36, 267] to measure the strength of modular organization in networks. Modularity can be defined as:

$$Q = \sum_{k=1}^{K} (e_{kk} - a_k^2)$$  \hspace{1cm} (C.1)

where $e_{kk}$ denotes the fraction of edges within the subgroup $k$ and $a_k$ is the fraction of the total edges present within the subgroup. If $\overline{d_k}$ is the average degree of individuals
in subgroup $k$, $\overline{d}_w^k$ is the average within subgroup degree and $s_k$ is the total number of individuals in the subgroup, then the equation C.1 can be rewritten as:

$$Q = \sum_{k=1}^{K} \left[ \overline{d}_w^k s_k - \left( \frac{\overline{d}_w^k s_k}{dn} \right)^2 \right] \quad (C.2)$$

where $\overline{d}$ is the average degree of the network, and the total number of individuals and subgroups in the network are $n$ and $K$, respectively.

Let $L_k$ be the total edges of individuals of a subgroup $k$ out of which $L_k^w$ are the edges present within the subgroup. Equation (2) can then be further reduced to:

$$Q = \sum_{k=1}^{K} \left[ \frac{L_k^w}{L} - \left( \frac{L_k}{L} \right)^2 \right] \quad (C.3)$$

where $L$ is the total edges present in the network.

We estimated Newman modularity, $Q$, for the largest connected component of each animal social network using the Louvain method as described in [149].

The highest possible modularity in a network is achieved when all individuals in a subgroup $k$ only interact with each other and no edges are present between subgroups (i.e., subgroups are disjointed). In other words, $Q_{max}$ of a network is when $L_k^w = L_k$. Equation S3 can be therefore written as

$$Q_{max} = \sum_{k=1}^{K} \left[ \frac{L_k}{L} - \left( \frac{L_k}{L} \right)^2 \right] \quad (C.4)$$

A network is more modular if its observed $Q$ value is closer to the maximum possible modularity. We thus normalize the strength of modular organization with respect to its $Q_{max}$ by computing relative modularity of networks as $Q_{rel} = \frac{Q}{Q_{max}}$. 

165
C.3 Mechanisms of modular organization

We test the role of several features of social organization in animal societies that may contribute towards the observed magnitude of relative modularity. These features include network size \( (n) \); (log of the) number of subgroups in the social network (network fragmentation); preferential association with own subgroup (subgroup cohesion, measured as the proportion of total contacts that occur within subgroups); variation in subgroup cohesion across the network; average and variation in subgroup sizes; average and variation in number of individual contacts (degree); and variation in contacts among subgroups (subgroup degree variation). We note that this is not an exhaustive list but includes factors that we believe can be important in the context of animal social networks. Average contacts of subgroups was not included in the analysis because mathematically it is equal to the average degree of individuals.

To examine the relative contribution of these factors towards network modularity, we ran the following mixed effects beta regression model using glmmADMB package (version 0.8.3.3) in R (version 3.2.3). All predictors were centered and scaled to unit variances to assign same prior importance to each predictor in the analysis [83]. The distributions of network size and number of subgroups were skewed to the right and therefore were natural log-transformed for further analyses. For all the variance predictors, we estimated coefficient of variation \( (CV = \sigma/\mu) \) to avoid possible correlation with their average predictor counterparts. To avoid multicollinearity, we estimated the variance inflation factor (VIF) implemented in the package ‘car’ in R, and removed predictors with VIF >5 [268]. We treated the species nested within the order of the social group and sociality as random effects in the model.
C.4 Generation of null networks

Two types of null networks were generated for each animal social network. Modular null networks were created in two steps: first, within-subgroup connections were randomized and the second step involved randomization of between-subgroup connections. Edges were randomized using the double-edge swap operation in the NetworkX package [79]. The modular null networks, therefore, had the same modular subdivision and degree sequence as the empirical networks, but were random with respect to other higher-order network properties. We generated homogeneous null networks by performing simultaneous edge swaps over within- and between-subgroup connections, which preserves the local contact heterogeneity but randomizes all higher-order features of the network including network modularity [185].

C.5 Generation of synthetic modular networks

We generated synthetic modular networks using modular random network generator described in [185]. This model allows generation of modular networks by varying the level of network fragmentation or subgroup cohesion while keeping other network properties (such as degree homophily, clustering coefficient) close to homogeneous null network with no modular structure. In this study we generated modular networks with 10,000 nodes. The strength of modular organization was varied by adjusting either subgroup cohesion or network fragmentation. As the total number of individuals was fixed across all generated synthetic networks, increasing network fragmentation (i.e., the number of subgroups) also coincided with decreasing the average subgroup size in the network. Unless otherwise noted, we assumed local contact heterogeneity in social groups to follow an exponential degree distribution with mean contact (degree) of 10.
C.6 Disease simulations

We performed Monte-Carlo simulations of discrete-time susceptible-infected-recovered (SIR) model of disease spread. Each disease simulation was initiated by infecting a randomly chosen individual in the social group. At subsequent time steps every infected individual(s) in the population could either transmit infection to a susceptible neighbor with probability parameter \( \beta \) or recover with probability parameter \( \gamma \). The disease simulation was terminated when there were no remaining infected nodes in the network. As we were interested only in major outbreaks, we considered only those simulations in our calculations where at least 10% individuals in the network acquired infection. Pathogen contagiousness was measured in terms of infection transmissibility. Transmissibility was defined as the probability of pathogen transmission from an infected to susceptible host during the period when the host is infectious. Assuming the individual’s recovery and infection transmission to be a Poisson process, transmissibility was calculated as: 

\[
T = \frac{\beta}{\beta + \gamma}. 
\]

**Disease simulations in Figure 2c and 5:** To compare epidemiological consequences of empirical networks, we simulated infectious disease spread with basic reproduction number, \( R_0 = 1.2 \). Basic reproduction number is defined as the average number of secondary infections caused by a single infected host in a completely susceptible population. Following ref. [269], we estimate transmissibility of pathogen corresponding as:

\[
T = R_0 \left( \frac{\langle k \rangle}{\langle k^2 \rangle - \langle k \rangle} \right) \tag{C.5} 
\]

where \( \langle k \rangle \) and \( \langle k^2 \rangle \) are the mean degree and mean square degree of the networks, respectively. Since empirical networks had varying values of \( \langle k \rangle \) and \( \langle k^2 \rangle \), we obtained a different value of pathogen transmissibility corresponding to a \( R_0 = 1.2 \) for each
empirical network. For consistent comparison, disease simulations on null networks were performed using the same value of pathogen transmissibility that was estimated for their empirical network counterpart using equation C.5.

### C.7 Threshold analysis

We identified the threshold of pathogen contagiousness below which there is a minimal risk of large outbreak in modular networks using the formula proposed in ref. [270]. Specifically, the threshold is estimated numerically as:

\[
\Delta = \frac{\sqrt{\langle \rho^2 \rangle - \langle \rho \rangle^2}}{\langle \rho \rangle}
\]

(C.6)

where \(\rho\) denotes the outbreak sizes of Monte-Carlo disease simulations, and \(\Delta\) is a variability measure. The epidemic threshold is estimated by performing disease simulations over a wide range of pathogen transmissibility and estimating the corresponding \(\Delta\) measure. The pathogen transmissibility where \(\Delta\) peaks is considered to be the epidemic threshold [270].
Table C.1: Multivariable analysis on determinants of modular organization in the dynamic networks of four species (*C. fellah*, *C. pennsylvanicus*, *P. lotor*, and *M. agrestis*).

<table>
<thead>
<tr>
<th>Explanatory variable</th>
<th>Effect size</th>
<th>95% confidence intervals</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>0.05</td>
<td>0.03 – 0.07</td>
</tr>
<tr>
<td>Network size</td>
<td>-0.03</td>
<td>-0.05 – 0.00</td>
</tr>
<tr>
<td>Network fragmentation*</td>
<td>0.48</td>
<td>0.44 – 0.52</td>
</tr>
<tr>
<td>Subgroup cohesion*</td>
<td>0.87</td>
<td>0.85 – 0.90</td>
</tr>
<tr>
<td>Subgroup cohesion variation</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Subgroup size average</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Subgroup size variation*</td>
<td>-0.10</td>
<td>-0.12 – -0.08</td>
</tr>
<tr>
<td>Individual degree average</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Individual degree variation</td>
<td>-0.02</td>
<td>-0.06 – 0.02</td>
</tr>
<tr>
<td>Subgroup degree variation*</td>
<td>-0.04</td>
<td>-0.07 – -0.02</td>
</tr>
<tr>
<td>Random effects</td>
<td></td>
<td>Variance estimate ($\sigma^2$)</td>
</tr>
<tr>
<td>Group identification</td>
<td></td>
<td>0.0002</td>
</tr>
</tbody>
</table>

Response variable is the relative modularity, $Q_{rel}$, of the social networks. Asterisks and bold text indicate significance. We treated the social group identity of each network as the random effect. Explanatory variables with VIF>5 were dropped from the model, and therefore their effect sizes therefore were not estimated.
Figure C.1: Empirical animal social networks and the corresponding homogeneous null networks included in Figure 2c. Animal social network abbreviations: AC = Acanthiza spp., CC = Cercopithecus campbelli, CCR = Crocuta crocuta, CF = Camponotus fellah, CP = Camponotus pennsylvanicus, BB = Bison bison, BL = Branta leucopsis, EP = Erythrocebus patas, GA = Gasterosteus aculeatus, HM = Haemorhous mexicanus, MA = Mirounga angustirostris, MAR = Macaca arctoides, MM = Macaca mulatta, MS = Myotis sodalis, OC = Ovis canadensis, PC = Papio cynocephalus, PL = Procyon lotor, PP = Pan paniscus, PT = Pan troglodytes, TR = Tiliqua rugosa, TT = Tursiops truncatus, ZA = Zonotrichia atricapilla. Numbers next to the species acronym indicate a separate social group.
Figure C.2: Threshold value of relative modularity above which the empirical networks tend to experience lower outbreak sizes as compared to homogeneous null networks in Figure 2c. The breakpoint is at $Q_{rel} \approx 0.618$. \( \Delta \) Outbreak size is the relative outbreak size and is calculated as $\frac{S_{null} - S_{emp}}{S_{emp}}$ where $S = \text{average outbreak size}$, $emp = \text{empirical network}$, $null = \text{homogeneous null networks}$. We recursively fit a piecewise linear mixed regression model with a sequence of increasing breakpoint values. The optimum breakpoint value was selected from the model with minimum deviance. To take into account any correlation within the taxonomic groups, we used the species nested with genus and taxonomic order of the groups as a random effect in the model. Sociality of the animal groups analyzed (relative solitary, social, and fission-fusion) was also included as a random effect.
Figure C.3: Extended figure of Figure 2c: comparisons between real (filled points) and their homogeneous null networks (tip of arrows) with respect to the percentage of infected individuals (outbreak size) due to an outbreak with basic reproduction number, $R_0$ equal to 1.45, 1.8, 2.4 and 4.8. Point color corresponds to the taxonomic class of the animal group. Social networks with non-significant modular subdivision (as indicated by $t$ test analysis) have been excluded. The generated homogeneous null networks preserve only the local heterogeneity of contacts among individuals; the arrows therefore indicate the change in direction and magnitude of outbreak size under the scenario where all higher-order structural complexities (including modular subdivisions) are removed from animal social networks.
Figure C.4: The values of network fragmentation and subgroup cohesion present in empirical animal social networks. Network fragmentation is the log-transformed value of the total subgroups present in the social network. Circles are the networks with $Q_{rel} < 0.618$ (the threshold value shown in Fig.2c of the main text and SI Appendix, Fig.S2) and triangles represent social networks above the modularity threshold. Animal social networks above the threshold experienced lower outbreak size as compared to homogeneous null networks in Fig.2C. In this study, we show that the disease consequences of modular subdivision in animal social networks are driven by network fragmentation and subgroup cohesion.
Figure C.5: Reduction in outbreak size experienced by empirical animal social networks as compared to their homogeneous null networks (as shown in Figure 2c) as a function of subgroup cohesion and (a) network fragmentation, and (b) subgroup size variation.
Figure C.6: Disease implications of modular subdivisions in synthetic modular networks as a function of subgroup cohesion and network fragmentation. Outbreak size and duration have been normalized to the maximum observed outbreak size and duration respectively. Subgroup cohesion measures the tendency of individuals to interact with members of their own subgroup, and is measured as the proportion of contacts within-subgroups to total contacts across the entire social network. In theory, a social network with two subgroups is considered to have $Q_{rel} = 0$ when the proportion of contacts within subgroups is equal to the proportion of contacts between subgroups (=0.5). The minimum subgroup cohesion realized in a network is therefore a function of the number of subgroups (fragmentation) present in the network. Circles, triangles and squares summarize disease outbreaks of pathogen transmissibility of 0.1, 0.18 and 0.3, respectively.
Figure C.7: (a) Outbreak size and (b) outbreak duration in networks with different heterogeneity in subgroup sizes. Constant = all subgroups have equal number of individuals; homogeneous = subgroup size distribution follows a Poisson distribution; heterogeneous = subgroup sizes follow an exponential distribution. For homogeneous and heterogeneous subgroup size distribution, the average subgroup size = $N$/network fragmentation, where $N$ is total number of individuals in the social group. Results summarize disease simulations of a pathogen with transmissibility = 0.1 over networks with 10 subgroups. Outbreak size and duration have been normalized to the maximum observed outbreak size and duration respectively.
Figure C.8: Disease burden in modular social networks with Poisson (purple bars) and exponential (orange bars) contact heterogeneity. The y-axis measures the change in outbreak size and duration of modular networks as compared to disease burden in homogeneous networks (with $Q_{rel}=0$), and is calculated as $O_{modular} - O_{non-modular}$, where $O_{modular}$ and $O_{non-modular}$ is the average outbreak size or duration of networks with $Q_{rel}>0$ and $Q_{rel}=0$, respectively. Errors bars represent standard error of the means.
Figure C.9: Percentage error in outbreak size predictions using modular and homogeneous null networks for 19 animal social networks due to an outbreak with $R_0 = 1.05, 1.45, 1.8$ and $2.0$. Percentage error is calculated as $(S_{emp} - S_{null})/S_{emp} \times 100$, where $S =$ outbreak size, $emp =$ empirical network, $null$ are modular or homogeneous null networks of the empirical network. The social networks are ordered according to the increasing value of relative modularity (red solid curve, secondary y-axis). The shaded region indicates the range of percentage error values below 15%. BA, *Brachyteles arachnoides*; BB, *Bison bison*; CC, *Cercopithecus campbelli*; CCR, *Crocuta crocuta*; CF, *C. fella*; CP, *C. pennsylvanicus*; DC, dairy cattle; DR, *Desmodus rotundus*; HM, *Haemorhous mexicanus*; MA, *M. angustirostris*; MF, *Macaca fuscata*; MM, *Macaca mulatta*; MT, *Macaca tonkeana*; PC, *Papio cynocephalus*; TR, *T. rugosa*; TT, *T. truncates*. Numbers denote separate groups of the same species.
Figure C.10: Outbreak size predictions from modular and homogeneous null networks as compared to the "true" outbreak size (empirical networks, gray vertical bars). Results are shown for simulations of a slowly ($R_0=0.85$) and rapidly ($R_0=4.8$) spreading infectious disease. Outbreak size is averaged across all disease simulations. Both the null networks (in most cases) yield identical and accurate outbreak size predictions.
Figure C.11: Highly modular networks are able to reduce outbreak size: empirical evidence from the spread of Babesia parasite in the field voles (Microtus agrestis). Babesia are spread by ticks and therefore networks based on common space use are appropriate to model the parasite transmission [40, 59]. We therefore used the bipartite networks described in [96] to examine the relationship between network modularity and parasite transmission. Specifically, generalized linear mixed model with binomial response distribution was used, where Babesia prevalence was entered as the response and the relative modularity of networks in the previous sampling period was entered as the explanatory variable (for details see [96]). The number of nodes and the total edges were also entered as explanatory variables in the model to control for variation in the sampling effort. Additionally, the site of data collection was included as a random effect. To control for autocorrelation, we included the parasite prevalence in the previous sampling period as an explanatory variable. Two models were fit - one that explored the relationship between parasite prevalence and relative modularity in networks where \( Q_{rel} \leq 0.6 \), and other that examined association between parasite prevalence and relative modularity for networks where \( Q_{rel} > 0.6 \). No significant association was found between network modularity and Babesia prevalence when \( Q_{rel} \leq 0.6 \) (\( \chi^2 = 0.21, P=0.65 \)). However, Babesia prevalence decreased with increasing modularity when \( Q_{rel} > 0.6 \) (\( \chi^2 = 12.50, P < 0.001 \)).
Figure C.12: Low modular networks do not limit global disease transmission: empirical evidence from the spread of a gastrointestinal parasite in brown spider monkeys (*Ateles hybridus*) [190]. Gastrointestinal parasites have been empirically shown to spread through the networks of physical interaction in brown spider monkeys [190]. In the social networks of brown spider monkeys described in ref. [190], we identified four connected subgroups with a $Q_{rel}$ of 0.24. Node colors indicate different subgroups. The triangle shaped nodes are the individuals that were reported infected with *Strongyloides sp.* in [190]. The parasite prevalence in the connected component of the network was 57%. No structural trapping was observed in the network (i.e., infections were reported in all subgroups), which suggests that the low modularity of spider monkey social network does not inhibit the global spread of parasites.
Figure C.13: Structural trapping of infectious disease increases with network fragmentation: empirical evidence from the spread of pneumonia in the big horn lambs (*Ovis canadensis*). Subgroups in big-horn lambs in [192] were identified based on direct or indirect association between animals to study the spread of pneumonia between individuals. This definition of subgroup implies that between-group interaction was minimal (i.e., high subgroup cohesion). Lamb mortality was considered to be a reasonable proxy of infection status [192]; we therefore calculated the outbreak size as the proportion of deaths within each lamb herd (population). In the figure, each population is represented by a different point color. We found negative (albeit weak) relationship between network fragmentation and outbreak size ($\chi^2 = 3.29$, $P = 0.069$).
Figure C.14: High subgroup cohesion can induce structural delay of infection spread: empirical evidence from the spread of *Mycoplasma gallisepticum* in house finches (*Haemorhous mexicanus*) [193]. The social network of the songbirds based on common feeder use has been associated with spread of *Mycoplasma gallisepticum* [193]. We estimated the relative modularity of the songbird social network to be 0.29 with three connected subgroups. Node colors indicate different subgroups. The triangle shaped nodes are the individuals that were reported infected with *Mycoplasma gallisepticum* in [193]. The infection prevalence in the connected component of network was 9%. No structural trapping was observed (i.e., infections were reported in all the subgroups) in the network, which suggests that low modularity does not inhibit global transmission. However, the local disease prevalence (b) was inversely correlated to subgroup cohesion in the interaction networks of songbirds. This suggests that high local (subgroup) cohesion induces structural delay of infection spread within the social network.
Figure C.15: Modular networks influence disease transmission of moderately, but not low, contagious pathogens/parasites. Networks based on physical proximity are associated with the spread of *Salmonella* in Australian sleepy lizards [197]. Our investigation of modular subdivisions in the social network revealed network modularity to be moderately high ($Q_{rel} = 0.624$). We used the transmission data of two Salmonella strains (published in ref. [197]) to validate our theoretical prediction that network modularity influences disease transmission of moderately, but not of low contagious pathogen. To do so, we first estimated pathogen transmissibility of the two *Salmonella* strains (Genotype 17 and 2) to be 0.043 ($R_0 = 0.48$) and 0.1788 ($R_0 = 2.02$), respectively. The estimated transmissibility value was used to perform simulations of *Salmonella* spread in the modular and homogeneous null networks of sleepy lizards. Finally, the outbreak size of disease simulations was compared to the real transmission data of the two genotypes in the sleepy lizard social network. For the low transmissible strain (Genotype 17), we found that the homogeneous and modular null networks produced identical and accurate outbreak size predictions. This suggests that modular subdivision does not influence transmission of low contagious parasites. Conversely, for the moderately transmissible strain (Genotype 2), modular null networks performed better in estimating the true outbreak size as compared to homogeneous null networks, which suggests that modular subdivision does influence the transmission of moderately transmissible parasites.
<table>
<thead>
<tr>
<th>Key findings of this study</th>
<th>Evidence</th>
<th>Specific findings</th>
<th>Notes</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Structural trapping by highly modular network localizes infection to a small proportion of subgroups</td>
<td>Plague (Yersinia pestis) infection in the great gerbil (Rhombomys opimus)</td>
<td>Plague spread easily within burrow systems (defined as subpopulations in the study) than between them. Probability that an animal was infectious or recovered was higher if its subpopulation was known to harbour infected animals in the recent past than if there was no evidence of recent infection.</td>
<td>Although network modularity was not explicitly calculated, limited movement of gerbil or infected fleas between burrow systems is indicative of a high network modularity.</td>
<td>[271]</td>
</tr>
<tr>
<td>Structural delay of infection spread</td>
<td>Canine distemper virus outbreak in Greater Yellowstone ecosystems wolves (Canis lupus) and coyotes (Canis Latrans) Epidemic of phocine distemper virus in the North Sea population of harbour seals (Phoca vitulina)</td>
<td>Low connectivity (indicative of high modularity) between host population increases outbreak duration When coupling between host patches is weak (indicative of high modularity), higher number of host patches (fragmentation) increases outbreak duration</td>
<td>Data driven model</td>
<td>[272]</td>
</tr>
<tr>
<td>Low levels of modular organization does not structurally trap infection</td>
<td>Escherichia coli transmission in wild elephants (Loxodonta africana)</td>
<td>No evidence that animals were more likely to be infected with E. coli from members of their own subgroup than members of other subgroups</td>
<td>Aggregation of animals in single large groups suggests social networks to be homogeneous (or low modular) during rainy periods. In addition, the authors observe high overlap in home-ranges around water sources, which indicates social networks are not highly modular during non-rainy seasons.</td>
<td>[274]</td>
</tr>
<tr>
<td>Effect of modular structure on disease transmission depends on pathogen contagiousness</td>
<td>Cryptosporidium spread in wild lemurs</td>
<td>Increasing network modularity reduces outbreak size for moderately transmissible pathogen, but not for low and high pathogen transmissibility</td>
<td>Pathogen transmissibility can be derived from Figure 7 based β and γ values. Data-driven model.</td>
<td>[275]</td>
</tr>
</tbody>
</table>
Appendix D

Chapter 5 Supplementary materials

Figure D.1: Average proportion of weak ties of individuals across the three social systems. Shown is the comparisons between networks where the weakest 5%, 10%, 15% and 20% of edges were removed. The boxplot shows the mean value inside the box, standard deviation as the length of the box, and the whiskers represent the minimum and maximum values.
Table D.1: Description of the network measures used in this study.

<table>
<thead>
<tr>
<th>Network measure</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Degree heterogeneity</td>
<td>Coefficient of variation (CV) in the degree distribution measured as the standard deviation in degree divided by the mean degree. A degree distribution with CV of 1 approximates an exponential distribution.</td>
</tr>
<tr>
<td>Degree homophily</td>
<td>The tendency of social partners to have a similar degree. Measured as the correlation coefficient between the degrees of neighboring nodes.</td>
</tr>
<tr>
<td>Global clustering coefficient</td>
<td>Propensity of social partners of a focal individual to interact with each other. Clustering coefficient of a node is measured as the fraction of all possible triangles through the node that exist in the network.</td>
</tr>
<tr>
<td>Weighted clustering coefficient</td>
<td>Global clustering coefficient of nodes in the social networks taking edge weights into account as described in [276].</td>
</tr>
<tr>
<td>Transitivity</td>
<td>Fraction of all possible triangles present in the interaction network.</td>
</tr>
<tr>
<td>Global betweenness centrality</td>
<td>Betweenness centrality is a measure of how central a node is in the network, and is defined as the number of shortest path that go through the focal node in the network. Nodes in an interaction network with high average betweenness centrality have a greater tendency to occupy socially central positions.</td>
</tr>
<tr>
<td>Weighted betweenness centrality</td>
<td>Average betweenness centrality of the sociality network taking edge weights into account.</td>
</tr>
<tr>
<td>Average group size</td>
<td>Average number of individuals present in groups within the interaction network, where size and composition of groups is determined by the Louvain algorithm.</td>
</tr>
<tr>
<td>Group cohesion</td>
<td>Proportion of the total interactions that occur within the social groups. High group cohesion indicates higher individual preferences to interact with members of own social group.</td>
</tr>
<tr>
<td>Network fragmentation</td>
<td>Log of the number of social groups present within the largest connected component of the interaction network, where the number of social groups is determined by the Louvain algorithm.</td>
</tr>
<tr>
<td>Relative modularity ($Q_{rel}$)</td>
<td>Normalized Newman modularity as described in [46]. Networks with higher $Q_{rel}$ tend to be highly fragmented and have high group cohesion.</td>
</tr>
<tr>
<td>Network diameter</td>
<td>The longest of the shortest path between any pair of nodes in the largest connected component of interaction network. Information typically spreads faster in networks with a smaller diameter.</td>
</tr>
<tr>
<td>Average individual degree</td>
<td>Average degree of individuals in the social group, where size and composition of groups in a social network is determined by the Louvain algorithm.</td>
</tr>
<tr>
<td>Average individual strength</td>
<td>Total edge weights of individuals, averaged over all individuals present in the social group. Average individual strength represents the amount of energy invested in social interaction by each individual in the network.</td>
</tr>
<tr>
<td>Average pairwise strength</td>
<td>Total edge weights of individuals, averaged over all pairwise interactions in the social network. Average pairwise represents the amount of energy invested in dyadic relationships by individuals in the social group.</td>
</tr>
<tr>
<td>Group connectivity</td>
<td>Average degree of individuals in the social group divided by total number of individuals present in the social group.</td>
</tr>
</tbody>
</table>

Note: All network metrics (with the exception of relative modularity, group cohesion and network fragmentation) were measured using the NetworkX package (https://networkx.github.io/) implemented in Python.
Figure D.2: Role of network structures in influencing disease transmission.
The three network measures shown are the ones included in the final model as shown in Table 1. Error bars represent 95% credible intervals. Credible intervals that do not include zero suggest significant association with disease transmission (red = significant effect, black = effect not significant).
Figure D.3: Disease costs of social organization due to interaction network structure. Disease cost has been quantified in terms of epidemic probability (likelihood of large outbreaks infecting at least 15% of individuals in the network), epidemic duration (time to epidemic extinction), and epidemic size (average percent of individuals infected in an epidemic outbreak), for different rates of pathogen transmissibility, $T$. Error bars represent standard errors, and different letters above the bars denote a significant difference between the means ($P < 0.05$).
Figure D.4: Disease implications of social systems due to pathogens with increasing transmissibility. Error bars represent standard errors, and different letters above the bars denote a significant difference between the means (P < 0.05). As expected, the probability and size of epidemics increase with increasing pathogen transmissibility. We note that the trends of epidemic probability and epidemic size are qualitatively similar, which is expected based on network epidemiology theory [277]. Duration of epidemics is similar across different values of pathogen transmissibility, except of lowly contagious pathogen. For low contagious pathogens, only a small proportion of simulations reach epidemic proportions (i.e., simulations that infect at least 15% of individuals in a population). The seemingly high epidemic duration in socially hierarchical species, and low epidemic duration in gregarious species when T < 0.15 is an artifact of low sample sizes.
Figure D.5: Role of network structures in influencing transmission of disease with an infectious period of 10 days ($\gamma = 0.1$). The three network measures shown are the ones included in the final model as shown in (Table 2). Error bars represent 95% credible intervals. Credible intervals that do not include zero suggest significant association with disease transmission (red = significant effect, black = effect not significant).
Figure D.6: Disease costs of social systems due to social network structure. The average infectious period of the simulated disease is 10 days ($\gamma=0.1$). Disease cost has been quantified in terms of epidemic probability (likelihood of large outbreaks infecting at least 15% of individuals in the network), epidemic duration (time to epidemic extinction), and epidemic size (average percent of individuals infected in an epidemic outbreak), for different rates of pathogen transmissibility, $T$. Error bars represent standard errors, and different letters above the bars denote a significant difference between the means ($P < 0.05$).
Figure D.7: Disease costs of the three social systems when ties less than 20% of highest edge weight are removed from the social networks. The infectious period of the simulated disease is 5 days ($\gamma = 0.2$). Disease cost has been quantified in terms of epidemic probability (likelihood of large outbreaks infecting at least 15% of individuals in the network), epidemic duration (time to epidemic extinction), and epidemic size (average percent of individuals infected in an epidemic outbreak), for different rates of pathogen transmissibility, $T$. Error bars represent standard errors, and different letters above the bars denote a significant difference between the means ($P < 0.05$).
Table D.2: Percentage of variance explained by the random effects included in the Bayesian generalized linear mixed model presented in Table 1 (assuming relative solitary species as the base social system).

<table>
<thead>
<tr>
<th>Random effect</th>
<th>Percent variance estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Phylogenetic correlation</td>
<td>78.84</td>
</tr>
<tr>
<td>Group identification</td>
<td>0.11</td>
</tr>
<tr>
<td>Interaction type</td>
<td>0.22</td>
</tr>
<tr>
<td>Edge type</td>
<td>0.24</td>
</tr>
<tr>
<td>Sampling scale</td>
<td>8.56</td>
</tr>
</tbody>
</table>

Table D.3: Results of cross validation using random sub-sampling method. All effect size estimates are within 95% credible intervals of effects estimated using the full data (Table 1).

<table>
<thead>
<tr>
<th>Degree heterogeneity</th>
<th>Focal</th>
<th>Base</th>
<th>Relatively solitary</th>
<th>Gregarious</th>
<th>Socially hierarchical</th>
</tr>
</thead>
<tbody>
<tr>
<td>Relatively solitary</td>
<td>-3.05 ± 0.65 &amp; -9.71 ± 1.53</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gregarious</td>
<td>-5.90 ± 1.02</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Socially hierarchical</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Degree homophily</th>
<th>Focal</th>
<th>Base</th>
<th>Relatively solitary</th>
<th>Gregarious</th>
<th>Socially hierarchical</th>
</tr>
</thead>
<tbody>
<tr>
<td>Relatively solitary</td>
<td>-0.14 ± 0.37</td>
<td>-3.60 ± 0.82</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gregarious</td>
<td>-3.16 ± 0.62</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Socially hierarchical</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Global betweenness centrality</th>
<th>Focal</th>
<th>Base</th>
<th>Relatively solitary</th>
<th>Gregarious</th>
<th>Socially hierarchical</th>
</tr>
</thead>
<tbody>
<tr>
<td>Relatively solitary</td>
<td>0.98 ± 0.71</td>
<td>-0.78 ± 0.90</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gregarious</td>
<td>-1.07 ± 0.70</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Socially hierarchical</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Global clustering coefficient</th>
<th>Focal</th>
<th>Base</th>
<th>Relatively solitary</th>
<th>Gregarious</th>
<th>Socially hierarchical</th>
</tr>
</thead>
<tbody>
<tr>
<td>Relatively solitary</td>
<td>0.71 ± 0.51</td>
<td>-4.59 ± 0.87</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gregarious</td>
<td>-4.94 ± 0.64</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Socially hierarchical</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Group cohesion</th>
<th>Focal</th>
<th>Base</th>
<th>Relatively solitary</th>
<th>Gregarious</th>
<th>Socially hierarchical</th>
</tr>
</thead>
<tbody>
<tr>
<td>Relatively solitary</td>
<td>0.32 ± 0.53</td>
<td>0.89 ± 0.84</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gregarious</td>
<td>0.35 ± 0.65</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Socially hierarchical</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Network fragmentation</th>
<th>Focal</th>
<th>Base</th>
<th>Relatively solitary</th>
<th>Gregarious</th>
<th>Socially hierarchical</th>
</tr>
</thead>
<tbody>
<tr>
<td>Relatively solitary</td>
<td>2.13 ± 0.68</td>
<td>-0.94 ± 1.00</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gregarious</td>
<td>-2.95 ± 0.76</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Socially hierarchical</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Network diameter</th>
<th>Focal</th>
<th>Base</th>
<th>Relatively solitary</th>
<th>Gregarious</th>
<th>Socially hierarchical</th>
</tr>
</thead>
<tbody>
<tr>
<td>Relatively solitary</td>
<td>-1.62 ± 1.00</td>
<td>3.47 ± 1.36</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gregarious</td>
<td>4.27 ± 1.14</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Socially hierarchical</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table D.4: Effect size estimates of two separate phylogenetically controlled Bayesian generalized liner mixed models examining the characteristics of social networks associated with (a) the type of interaction recorded and (b) the spatial scale of data collection.

<table>
<thead>
<tr>
<th>Network metric</th>
<th>Interaction type (base = association)</th>
<th>Sampling scale (base = social sampling)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Degree heterogeneity</td>
<td>-0.34 [-2.36, 1.59]</td>
<td>-0.56 [-3.41, 2.41]</td>
</tr>
<tr>
<td>Degree homophily</td>
<td>-0.23 [-0.92, 0.45]</td>
<td>0.61 [-0.57, 1.79]</td>
</tr>
<tr>
<td>Global betweenness centrality</td>
<td>0.14 [-0.47, 0.77]</td>
<td>-3.97 [-7.68, -0.36]</td>
</tr>
<tr>
<td>Global clustering coefficient</td>
<td>-0.52 [-1.53, 0.46]</td>
<td>-2.89 [-5.08, -0.72]</td>
</tr>
<tr>
<td>Group cohesion</td>
<td>0.06 [-0.91, 1.08]</td>
<td>1.35 [-0.38, 3.09]</td>
</tr>
<tr>
<td>Network fragmentation</td>
<td>0.52 [-1.10, 2.11]</td>
<td>0.40 [-1.85, 2.62]</td>
</tr>
<tr>
<td>Network diameter</td>
<td>0.30 [-1.37, 1.89]</td>
<td>-2.85 [-6.02, 0.09]</td>
</tr>
</tbody>
</table>

The predictors included are identical to the ones described in Table 1. Random effect for model (a) included - animal species, group identification, edge weight type (unweighted vs. weighted), social system (relatively solitary, gregarious or socially hierarchical) and sampling scale (social sampling vs. spatial sampling). For model (b), the random effect of sampling scale was replaced with interaction type (association vs. interaction). Shown are the posterior means of the expected change in log-odds of being associated with (a) networks where interactions were recorded between individuals as opposed to associations, and (b) networks that were collected on a population scale rather than local groups, with one-unit increase in the network measure. The 95% credible intervals are included in brackets. Significant terms with pMCMC < 0.05 are indicated in bold, where pMCMC is the proportion of MCMC samples that cross zero.
Table D.5: Filtered networks with 5% weakest edges removed: Effect size estimates of the generalized linear mixed models (by MCMCglmm) examining the characteristics of filtered interaction network structure for different species sociality.

<table>
<thead>
<tr>
<th>Degree heterogeneity</th>
<th>Focal</th>
<th>Relatively solitary</th>
<th>Gregarious</th>
<th>Socially hierarchical</th>
</tr>
</thead>
<tbody>
<tr>
<td>Relatively solitary</td>
<td>-7.81 [-15.86, -0.73]</td>
<td>-9.63 [-17.53, -2.10]</td>
<td>-1.68 [-4.76, 1.43]</td>
<td></td>
</tr>
<tr>
<td>Gregarious</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Socially hierarchical</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Degree homophily</th>
<th>Focal</th>
<th>Relatively solitary</th>
<th>Gregarious</th>
<th>Socially hierarchical</th>
</tr>
</thead>
<tbody>
<tr>
<td>Relatively solitary</td>
<td>-0.51 [-2.14, 3.24]</td>
<td>-0.78 [-3.77, 2.16]</td>
<td>-1.16 [-2.70, 0.33]</td>
<td></td>
</tr>
<tr>
<td>Gregarious</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Socially hierarchical</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Global betweenness centrality</th>
<th>Focal</th>
<th>Relatively solitary</th>
<th>Gregarious</th>
<th>Socially hierarchical</th>
</tr>
</thead>
<tbody>
<tr>
<td>Relatively solitary</td>
<td>3.73 [-2.65, 10.92]</td>
<td>3.41 [-2.92, 10.72]</td>
<td>-0.16 [-1.70, 1.60]</td>
<td></td>
</tr>
<tr>
<td>Gregarious</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Socially hierarchical</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Global clustering coefficient</th>
<th>Focal</th>
<th>Relatively solitary</th>
<th>Gregarious</th>
<th>Socially hierarchical</th>
</tr>
</thead>
<tbody>
<tr>
<td>Relatively solitary</td>
<td>0.45 [-3.52, 4.59]</td>
<td>-2.77 [-7.01, 1.51]</td>
<td>-2.65 [-4.45, -0.81]</td>
<td></td>
</tr>
<tr>
<td>Gregarious</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Socially hierarchical</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Group cohesion</th>
<th>Focal</th>
<th>Relatively solitary</th>
<th>Gregarious</th>
<th>Socially hierarchical</th>
</tr>
</thead>
<tbody>
<tr>
<td>Relatively solitary</td>
<td>0.74 [-2.62, 4.14]</td>
<td>1.48 [-2.16, 5.08]</td>
<td>0.91 [-0.64, 2.46]</td>
<td></td>
</tr>
<tr>
<td>Gregarious</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Socially hierarchical</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Network fragmentation</th>
<th>Focal</th>
<th>Relatively solitary</th>
<th>Gregarious</th>
<th>Socially hierarchical</th>
</tr>
</thead>
<tbody>
<tr>
<td>Relatively solitary</td>
<td>5.10 [0.23, 10.35]</td>
<td>2.03 [-2.67, 6.86]</td>
<td>-2.14 [-4.43, 0.03]</td>
<td></td>
</tr>
<tr>
<td>Gregarious</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Socially hierarchical</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Network diameter</th>
<th>Focal</th>
<th>Relatively solitary</th>
<th>Gregarious</th>
<th>Socially hierarchical</th>
</tr>
</thead>
<tbody>
<tr>
<td>Relatively solitary</td>
<td>-1.94 [-6.06, 2.18]</td>
<td>-1.49 [-6.09, 3.33]</td>
<td>0.31 [-2.36, 3.11]</td>
<td></td>
</tr>
<tr>
<td>Gregarious</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Socially hierarchical</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The filtered interaction networks were created by removing edges with edge-weights less than 5th percentile of edge-weight distribution in the network. Shown are the posterior means of the expected change in log-odds of being in focal social system (column headers), as compared to the base social system (row headers), with one-unit increase in the network measure. The 95% credible intervals are included in brackets. Significant terms with pMCMC < 0.05 are indicated in bold, where pMCMC is the proportion of MCMC samples that cross zero.
Table D.6: Filtered networks with 10% weakest edges removed: Effect size estimates of the generalized linear mixed models (by MCMCglmm) examining the characteristics of filtered interaction network structure for different species sociality.

<table>
<thead>
<tr>
<th></th>
<th>Degree heterogeneity</th>
<th>Degree homophily</th>
<th>Global betweenness centrality</th>
<th>Global clustering coefficient</th>
<th>Group cohesion</th>
<th>Network fragmentation</th>
<th>Network diameter</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Focal</td>
<td>Relatively solitary</td>
<td>Gregarious</td>
<td>Socially hierarchical</td>
<td>Focal</td>
<td>Relatively solitary</td>
<td>Gregarious</td>
</tr>
<tr>
<td>Relatively solitary</td>
<td>-0.98 [-4.41, 2.30]</td>
<td>-0.14 [-3.21, 3.05]</td>
<td>1.01 [-0.72, 2.83]</td>
<td>0.58 [-2.41, 3.70]</td>
<td>-1.31 [-4.19, 1.30]</td>
<td>2.45 [-1.60, 6.51]</td>
<td>-1.19 [-4.59, 1.99]</td>
</tr>
<tr>
<td>Gregarious</td>
<td></td>
<td></td>
<td>0.87 [-1.28, 2.98]</td>
<td>-0.51 [-1.73, 0.71]</td>
<td></td>
<td>1.12 [-0.43, 2.54]</td>
<td></td>
</tr>
<tr>
<td>Socially hierarchical</td>
<td></td>
<td></td>
<td></td>
<td>-0.68 [-3.75, 2.42]</td>
<td></td>
<td>-0.36 [-3.29, 2.45]</td>
<td></td>
</tr>
</tbody>
</table>

The filtered interaction networks were created by removing edges with edge-weights less than 10th percentile of edge-weight distribution in the network. Shown are the posterior means of the expected change in log-odds of being in focal social system (column headers), as compared to the base social system (row headers), with one-unit increase in the network measure. The 95% credible intervals are included in brackets. Significant terms with pMCMC < 0.05 are indicated in bold, where pMCMC is the proportion of MCMC samples that cross zero.
Table D.7: Filtered networks with 15% weakest edges removed: Effect size estimates of the generalized linear mixed models (by MCMCglmm) examining the characteristics of filtered interaction network structure for different species sociality.

<table>
<thead>
<tr>
<th></th>
<th>Focal</th>
<th>Relatively solitary</th>
<th>Gregarious</th>
<th>Socially hierarchical</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Degree heterogeneity</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Base</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Relatively solitary</td>
<td>1.04</td>
<td>[-1.90, 4.02]</td>
<td>1.02</td>
<td>[-1.78, 4.04]</td>
</tr>
<tr>
<td>Gregarious</td>
<td>0.12</td>
<td>[-1.72, 2.16 ]</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Degree homophily</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Base</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Relatively solitary</td>
<td>0.06</td>
<td>[-1.80, 1.94]</td>
<td>-0.95</td>
<td>[-3.02, 1.07]</td>
</tr>
<tr>
<td>Gregarious</td>
<td>-0.88</td>
<td>[-2.07, 0.37 ]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Socially hierarchical</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Global betweenness centrality</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Base</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Relatively solitary</td>
<td>6.49</td>
<td>[0.50, 12.89]</td>
<td>6.45</td>
<td>[0.55, 13.02]</td>
</tr>
<tr>
<td>Gregarious</td>
<td>0.03</td>
<td>[-1.06, 1.24 ]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Socially hierarchical</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Global clustering coefficient</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Base</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Relatively solitary</td>
<td>0.67</td>
<td>[-1.95, 3.52]</td>
<td>-0.29</td>
<td>[-3.20, 2.54]</td>
</tr>
<tr>
<td>Gregarious</td>
<td>-0.64</td>
<td>[-1.94, 0.61]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Socially hierarchical</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Group cohesion</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Base</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Relatively solitary</td>
<td>-1.24</td>
<td>[-4.21, 1.97]</td>
<td>-0.03</td>
<td>[-3.19, 3.33]</td>
</tr>
<tr>
<td>Gregarious</td>
<td>1.21</td>
<td>[-0.28, 2.69 ]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Socially hierarchical</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Network fragmentation</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Base</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Relatively solitary</td>
<td>0.38</td>
<td>[-3.37, 4.07]</td>
<td>0.31</td>
<td>[-3.51, 4.24]</td>
</tr>
<tr>
<td>Gregarious</td>
<td>-0.25</td>
<td>[-2.27, 1.68]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Socially hierarchical</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Network diameter</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Base</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Relatively solitary</td>
<td>-0.31</td>
<td>[-3.44, 2.60]</td>
<td>-0.69</td>
<td>[-3.70, 2.49]</td>
</tr>
<tr>
<td>Gregarious</td>
<td>-0.23</td>
<td>[-1.82, 1.58]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Socially hierarchical</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The filtered interaction networks were created by removing edges with edge-weights less than 15th percentile of edge-weight distribution in the network. Shown are the posterior means of the expected change in log-odds of being in focal social system (column headers), as compared to the base social system (row headers), with one-unit increase in the network measure. The 95% credible intervals are included in brackets. Significant terms with pMCMC < 0.05 are indicated in bold, where pMCMC is the proportion of MCMC samples that cross zero.
Table D.8: Filtered networks with 20% weakest edges removed: Effect size estimates of the generalized linear mixed models (by MCMCglmm) examining the characteristics of filtered interaction network structure for different species sociality.

<table>
<thead>
<tr>
<th></th>
<th>Focal</th>
<th>Relatively solitary</th>
<th>Gregarious</th>
<th>Socially hierarchical</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Degree heterogeneity</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Base</td>
<td>Relatively solitary</td>
<td>1.73 [-1.58, 5.07]</td>
<td>1.77 [-1.47, 5.09]</td>
<td>0.30 [-1.76, 2.41]</td>
</tr>
<tr>
<td></td>
<td>Gregarious</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Socially hierarchical</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Degree homophily</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Base</td>
<td>Relatively solitary</td>
<td>-0.18 [-2.07, 1.71]</td>
<td>-1.46 [-3.52, 0.59]</td>
<td>-1.16 [-2.42, 0.03]</td>
</tr>
<tr>
<td></td>
<td>Gregarious</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Socially hierarchical</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Global betweenness centrality</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Base</td>
<td>Relatively solitary</td>
<td>5.47 [-0.41, 11.77]</td>
<td>5.70 [-0.29, 11.97]</td>
<td>0.21 [-1.01, 1.31]</td>
</tr>
<tr>
<td></td>
<td>Gregarious</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Socially hierarchical</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Global clustering coefficient</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Base</td>
<td>Relatively solitary</td>
<td>0.79 [-1.85, 3.22]</td>
<td>-0.45 [-3.03, 2.16]</td>
<td>-0.94 [-2.17, 0.32]</td>
</tr>
<tr>
<td></td>
<td>Gregarious</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Socially hierarchical</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Group cohesion</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Base</td>
<td>Relatively solitary</td>
<td>-0.44 [-3.32, 2.56]</td>
<td>0.26 [-3.04, 3.28]</td>
<td>0.63 [-0.93, 2.19]</td>
</tr>
<tr>
<td></td>
<td>Gregarious</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Socially hierarchical</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Network fragmentation</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Base</td>
<td>Relatively solitary</td>
<td>-0.54 [-4.85, 3.40]</td>
<td>-1.73 [-6.23, 2.45]</td>
<td>-1.53 [-3.67, 0.69]</td>
</tr>
<tr>
<td></td>
<td>Gregarious</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Socially hierarchical</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Network diameter</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Base</td>
<td>Relatively solitary</td>
<td>-0.14 [-3.12, 3.17]</td>
<td>-0.29 [-3.57, 3.81]</td>
<td>0.14 [-1.62, 2.06]</td>
</tr>
<tr>
<td></td>
<td>Gregarious</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Socially hierarchical</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The filtered interaction networks were created by removing edges with edge-weights less than 20th percentile of edge-weight distribution in the network. Shown are the posterior means of the expected change in log-odds of being in focal social system (column headers), as compared to the base social system (row headers), with one-unit increase in the network measure. The 95% credible intervals are included in brackets. Significant terms with pMCMC < 0.05 are indicated in bold, where pMCMC is the proportion of MCMC samples that cross zero.
Appendix E

Chapter 6 Supplementary materials

Figure E.1: Robustness plot of (a) INoDS, (b) $k$-test and (c) network position test to three common forms of missing data - missing nodes, missing edges and missing infection cases. Null expectation in INoDS and network position test was generated by permuting the edge connections of the observed networks, creating an ensemble of null networks. In $k$-test, the location of infection cases within the observed network are permuted, creating a permuted distribution of $k$-statistic [201].
Figure E.2: Estimation of the social (\(\beta\)) and the asocial (\(\alpha\)) transmission parameter by INoDS under three forms of missing data conditions - (a) missing nodes, (b) missing edges and (c) missing infection cases. Simulations of susceptible-infected (SI) model of infectious disease spread were performed on static network with 100 nodes, Poisson degree distribution, and an average degree of 3. Each boxplot summarizes the results of 10 independent disease simulations; the horizontal line in the middle is the mean of estimated parameter values, the top and the bottom horizontal line is the standard deviation, and the tip of the vertical line represents the maximum/minimum value. The solid red line represents the true value of \(\beta\) used in the disease simulations. Since the simulations were performed on a known synthetic network, the expected value of asocial transmission parameter is zero.

Figure E.3: Relative error in the estimations of parameter \(\beta\) under missing data conditions with and without the inclusion of the asocial parameter (\(\alpha\)) in the INoDS formulation. The simulated infectious disease spread (SI model, pathogen transmissibility = 0.03) was performed on static network with 100 nodes, Poisson degree distribution, and a average degree of 3. Relative error was calculated as \(\frac{\beta - \beta'}{\beta}\), where \(\beta\) is the transmissibility of the simulated pathogen (=0.03) and \(\beta'\) is the value of social transmission parameter estimated.
Figure E.4: Relationship between social and asocial force of infection with increasing percentage of missing data. Each boxplot summarizes the results of 10 independent disease simulations (indicated by points); the horizontal line in the middle is the average percent transmission events where the asocial force is greater than the infection force contributed by the social connections. The top and the bottom horizontal line is the standard deviation, and the tip of the vertical line represents the maximum/minimum value.
Figure E.5: Plot of sensitivity and specificity of (a) INoDS, (b) $k$-test and (c) network position test to three common forms of missing data - missing nodes, missing edges and missing infection cases. The observed network in this case is a static network with 100 nodes, Poisson degree distribution and an average network degree of 3. Simulations of pathogen spread with transmissibility = 0.03 were performed through the observed static network. Null expectation in INoDS and network position test was generated by permuting the edge connections of the observed networks, creating an ensemble of null networks. In $k$-test, the location of infection cases within the observed network are permuted, creating a permuted distribution of $k$-statistic [201]. All P-values obtained for significance testing were corrected for multiple comparisons.
Figure E.6: Identifying the contact network model of *Crithidia* spread in bumble bee colony (colony UN1). Edges in the contact network models represent physical interaction between the bees. Since the networks were fully connected, a series of filtered contact networks were constructed by removing weak weighted edges in the network. The x-axis represents the edge-weight threshold that was used to removed weak edges in the network. Two types of edge weights were tested - (a) duration and (b) frequency of contacts. In addition, across all ranges of percent weak edges removed, the two types of weighted network were converted to (c-d) binary networks. The results shown are estimated values of social transmission parameter $\beta$, and estimated values of asocial transmission parameter $\alpha$, for the different contact network hypotheses. The faded bars correspond to networks where the asocial force was higher than the social force for all transmission events (which indicates that the network hypothesis does not explain the spread of *Crithidia* infection). Numbers above bars indicate the log Bayesian (marginal) evidence of the networks that were detected to have statistically significant higher predictive predictive power as compared to an ensemble of null networks ($P < 0.05$, corrected for multiple comparisons). We note that frequency networks with more than 25% weak edge removed failed to converge in (c) and (d), and therefore the transmission parameter associated with these contact networks were not estimated.
Bibliography


215


[113] Danna Hinderle, Rebecca L. Lewison, Andrew D. Walde, Doug Deutschman, and William I. Boarman. The effects of homing and movement behaviors


[193] James S Adelman, Sahnzi C Moyers, Damien R Farine, and Dana M Hawley. Feeder use predicts both acquisition and transmission of a contagious pathogen


232


