bioRxiv preprint doi: https://doi.org/10.1101/2022.07.28.501936; this version posted August 1, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

The spectral underpinnings of pathogen spread on animal networks

Authors: Nicholas M. Fountain-Jones[1*], Mathew Silk[2,3], Raima Appaw[1], Rodrigo Hamede[1], Julie Rushmore[4], Kimberly VanderWaal[5], Meggan Craft[6], Scott Carver[1] & Michael Charleston[1]

Affiliations:

[1] School of Natural Sciences, University of Tasmania, Hobart Australia 7001.

[2]CEFE, University of Montpellier, CNRS, EPHE, IRD, University of Paul Valéry Montpellier 3, Montpellier, France

[3]Centre for Ecology and Conservation, University of Exeter, Penryn Campus, Penryn, UK

[4] Odum School of Ecology, University of Georgia, Athens, GA USA.

[5] Department of Veterinary Population Medicine, University of Minnesota, St Paul, MN, USA

[6] Department of Ecology, Evolution, and Behavior, University of Minnesota, St Paul, MN, USA

*Nick.FountainJones@utas.edu.au

Abstract

Predicting what factors promote or protect populations from infectious disease is a fundamental epidemiological challenge. Social networks, where nodes represent hosts and edges represent direct or indirect contacts between them, are key to quantifying these aspects of infectious disease dynamics. However, understanding the complex relationships between network structure and epidemic parameters in predicting spread has been out of reach. Here we draw on advances in spectral graph theory and interpretable machine learning, to build predictive models of pathogen spread on a large collection of empirical networks from across the animal kingdom. Using a small set of network spectral properties, we were able to predict pathogen spread with remarkable accuracy for a wide range of transmissibility and recovery rates. We validate our findings using well studied host-pathogen systems and provide a

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.28.501936; this version posted August 1, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

flexible framework for animal health practitioners to assess the vulnerability of a particular network to pathogen spread.

Introduction

Capturing patterns of direct or indirect contacts between hosts is crucial to model pathogen spread in populations (Newman 2002; Craft 2015; Sah *et al.* 2018, 2021). Increasingly, contact network approaches, where hosts are nodes and edges reflect interactions between hosts, play a central role in epidemiology and disease ecology (e.g., Meyers *et al.* 2005; Bansal *et al.* 2007; Eames *et al.* 2015; White *et al.* 2017). Incorporating networks allows models to capture the heterogeneity of contacts between individuals that can provide more nuanced and reliable estimates of pathogen spread, including in wildlife populations (e.g., Meyers *et al.* 2006; Bansal *et al.* 2010; Craft *et al.* 2011). Formulating general rules for how easy-to-calculate network structure properties may promote or restrict pathogen spread can reveal important insights into how host behaviour can mediate epidemic outcomes (Sah *et al.* 2017), and provide practitioners with a proxy for how vulnerable a population is to disease without extensive simulations (Silk *et al.* 2017; Sah *et al.* 2018). Further, network structural properties can be incorporated into traditional susceptible–infected–recovered (SIR) models to account for contact heterogeneity when predicting pathogen dynamics across populations (e.g., Meyers *et al.* 2005; Bansal *et al.* 2007).

However, it remains unclear whether one structural characteristic or a combination of characteristics can reliably predict pathogen dynamics across systems (Ames *et al.* 2011; Sah *et al.* 2018). For example, species that are more social tend to have more clustered or “modular” networks, and this modularity has been found to increase (Lentz *et al.* 2012), reduce (Salathé & Jones 2010) or have little effect (Sah *et al.* 2018) on outbreak size across

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.28.501936; this version posted August 1, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

different biological systems. The average number of contacts between hosts can be identical across networks and yet still result in substantially different outbreak patterns (Ames *et al.* 2011). Even the apparent size of the network, often constrained by limitations of sampling, can impact estimates of pathogen spread, particularly in wildlife populations (McCabe & Nunn 2018). As network characteristics, such as network size and modularity, are often correlated (Newman 2006; Silk *et al.* 2017) and can have complex impacts on spread (Sah *et al.* 2017; McCabe & Nunn 2018; Porter 2020), determining network characteristics that promote large outbreaks, for example, remains a fundamental question in infectious disease biology (Sah *et al.* 2018).

Searching for general relationships between network structure and pathogen spread in animal populations is further challenged, as the relationship is also affected by pathogen traits, such as infectiousness and recovery rate. For example, modularity appears to make no difference to disease outcomes for highly infectious pathogens (Sah *et al.* 2017). Diseases with long recovery rates can increase outbreak size across networks as well (Shu *et al.* 2016). Given that we rarely have reliable estimates of pathogen traits in wild populations (e.g., for different probabilities of infection per contact, or recovery rates) anyway, any predictive model of the relationship between spread and network structure would ideally be generalizable across pathogens.

Advances in spectral graph theory offer an additional set of measures based on the **spectrum** of a network rather than average node or edge level attributes. A graph spectrum is the set of **eigenvalues** (often denoted with a Greek lambda λ) of a matrix representation of a network (see Text Box 1 for further definitions for terminology in bold). Theoretical studies have shown relationships between particular eigenvalues and connectivity across networks are

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.28.501936; this version posted August 1, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

independent of pathogen propagation models (Prakash *et al.* 2010). For example, networks with a high **Fiedler value** (the second smallest eigenvalue of the network's **Laplacian matrix**) are "more connected" than those with low values. It has been found that, in ecological networks for example, if the Fiedler value is sufficiently large, removing edges will have little effect on overall network connectivity (Kumar *et al.* 2019), but whether this lack of effect is mirrored by pathogen dynamics is not yet clear. Another quantity of interest is **spectral radius –** the largest absolute value of the eigenvalues of its **adjacency matrix**. The link between the spectral radius and epidemiological dynamics is better understood, with theoretical work showing that this value closely mirrors both epidemic behaviour and network connectivity (Prakash *et al.* 2010) and has been used to understand vulnerability of cattle networks to disease (Darbon *et al.* 2018). For example, networks with the same number of edges and nodes but higher spectral radius (λ_1) are more vulnerable to outbreaks than networks with low spectral radius ($\lambda_1 \rightarrow 1$). We hypothesize that spectral measures such as these have great potential to improve our ability to predict dynamics of pathogen spread on networks, where previous methods such as modularity have proved inadequate (Sah *et al.* 2017).

We assess the predictive capability of spectral values compared to other structural attributes such as **modularity** (Newmans' Q; Newman 2006)) using advances in machine learning to construct non-linear models of simulated pathogen spread across a large collection of empirical animal networks including those from the Animal Social Network Repository (ASNR) (Sah *et al.* 2019). The ASNR is a large repository of empirical contact networks that provides novel opportunities to test the utility of spectral values in predicting spread across a wide variety of, mostly animal, taxa across a spectrum of social systems -- from eusocial ants (Arthropoda: Formicidae) to more solitary species such as the desert tortoise (*Gopherus*

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.28.501936; this version posted August 1, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

agassizii). Farmed domestic animals were not included in our analyses. We combined networks from this resource with other published networks, including badgers (*Meles meles*) (Weber *et al.* 2013), giraffes (*Giraffa camelopardalis*) (VanderWaal *et al.* 2014) and chimpanzees (*Pan troglodytes*) (Rushmore *et al.* 2013) to generate a dataset of over 600 unweighted networks from 51 species. We then simulated pathogen spread using a variety of SIR parameters and harnessed recent advances in multivariate interpretable machine learning models (*MrIML;* (Fountain-Jones *et al.* 2021)) to construct predictive models across SIR parameter space. As many species were represented by multiple networks, often over different populations and or timepoints and constructed in different ways (e.g., some edges reflected spatial proximity rather than direct contact), we included species and network construction variables in our models to account for these correlations in addition to exploring the diversity of network structures across the animal kingdom. Our interpretable machine learning models identify putative threshold values for the vulnerability of a network to pathogen spread that can be used by practitioners to understand outbreak risk across systems.

We test how well our network structure estimates of pathogen spread, trained on SIR simulation results, generalize to more complex pathogen dynamics in the wild We utilize two well studied wildlife-pathogen systems to assess how our predictions compare to empirical estimates of spread; *Mycobacterium bovis* (the bacterium that causes bovine tuberculosis (bTB)) in badger populations and devil facial tumour disease (DFTD) in Tasmanian devil (*Sarcophilus harrisii*) populations (Hamede *et al.* 2009). We demonstrate that using spectral measures of network structure alone can provide a useful proxy for disease vulnerability with estimates of prevalence comparable to those empirically derived. Further, we provide a user-friendly app that utilizes our models to provide practitioners with predictions, for example, of the prevalence of a pathogen across a variety of spread scenarios

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.28.501936; this version posted August 1, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

using a user-supplied network without the need for lengthy simulation. The url for this "Shiny" app is https://spreadpredictr.shinyapps.io/spreadpredictr/.

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.28.501936; this version posted August 1, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Text Box 1: Terminology used in this paper.

A *graph* (or "*network*") is a collection of *nodes* and a collection of *edges* connecting the nodes in pairs, e.g., nodes x, y joined by edge (x,y). We define the *size* of the network – usually n, as the number of nodes (this usage differs from other strict mathematical definitions, but we feel this is more intuitive). Two nodes are said to be *adjacent* if they are connected by an edge, and the number of vertices adjacent to a given vertex x is called its *degree*, $\deg(x)$. Edges may be directed, in which case edge (x,y) is different from edge (y,x), but in our analyses we treat them as *undirected*, so $(x,y)=(y,x)$. Graphs can be represented naturally by matrices whose rows and columns are indexed by the nodes $(1,2,\ldots,n)$: the obvious one is the *adjacency matrix* A, whose (i,j)-th entry A_{ij} is 1 if nodes i and j are adjacent, and 0 otherwise. A is symmetric and $n \times n$, as are all the matrices in this work. Another useful matrix is the *degree* matrix D, in which D_{ij} is the degree of node i if $i=j$, and 0 otherwise. The *Laplacian* matrix L is the most complex one we use herein, but is easily calculated using $L_{ij} = D_{ij} – A_{ij}$.

The *eigenvalues* of a matrix are solutions to the matrix equation $M\boldsymbol{v} = \lambda\boldsymbol{v}$, where M is a matrix and $\boldsymbol{v}$ a vector of the appropriate size. Solving for $\boldsymbol{v}$ yields λ. These eigenvalues, ordered by their size, form the *spectrum* of a graph, as derived using any of the matrices just described. The *Fiedler value* of a graph is the second-smallest eigenvalue of L, and the *spectral radius* is the largest eigenvalue of A.

Measures of *Modularity* such as the Newman Q coefficient capture the strength of division within a network by quantifying the density of edges within and between subgroups. When there is no division within the network as the density of edges is the same between and within subgroups $Q = 0$, whereas higher values of Q indicate stronger divisions (Newman 2006). As Q scales with network size (small networks being generally less modular), relative

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.28.501936; this version posted August 1, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

modularity (Q_{rel}) allows for comparison across network sizes by normalizing Q using the maximum possible modularity for the network (Q_{max}) (Sah *et al*. 2017).

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.28.501936; this version posted August 1, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Results

Diversity of network structures

We identified substantial variation in network structure across animal taxa. The static unweighted animal social networks in our database ranged from nearly completely unconnected (Spectral radius $\lambda_1 \sim 1$, Fiedler value ~0, not included in our predictive models) to highly connected (Spectral radius $\lambda_1 \sim 160$, Fiedler value ~ 140, Fig. 1). Similarly, the networks ranged from homogeneous (i.e., not modular, Q_{rel}□=□0, see the Text box for a definition) to highly modular and subdivided (Q_{rel}□> 0.8,). Our principal component analysis (PCA) identified key axes of structural variation across empirical networks (Fig. 2). The first principal component (PC1) distinguished networks that had a large diameter and mean path length and were highly modular (negative values), from networks with a high mean degree and transitivity (positive values, Fig. 2, see Table S2). The second principal component (PC2) separated networks based on network size (number of nodes), maximum degree and the network duration (i.e., the time period over which the network data was collected, Fig. 2). The eusocial ant networks (*Camponotus fellah*, Insecta: Hymenoptera) and mammal networks tended to cluster separately (Fig. 1), with the other taxonomic classes dispersed between these groups (Fig. 1) or species (see Fig. S1 for clustering by species). The networks' spectral properties (the Fiedler value and spectral radius) explained a unique portion of structural variance that did not covary with other variables (see Table S1 for vector loadings and Fig S2 for all pair-wise correlations). We found variables such as mean degree and transitivity the most correlated with the other variables and were excluded from further analysis (Tables S2, Fig S2).

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.28.501936; this version posted August 1, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

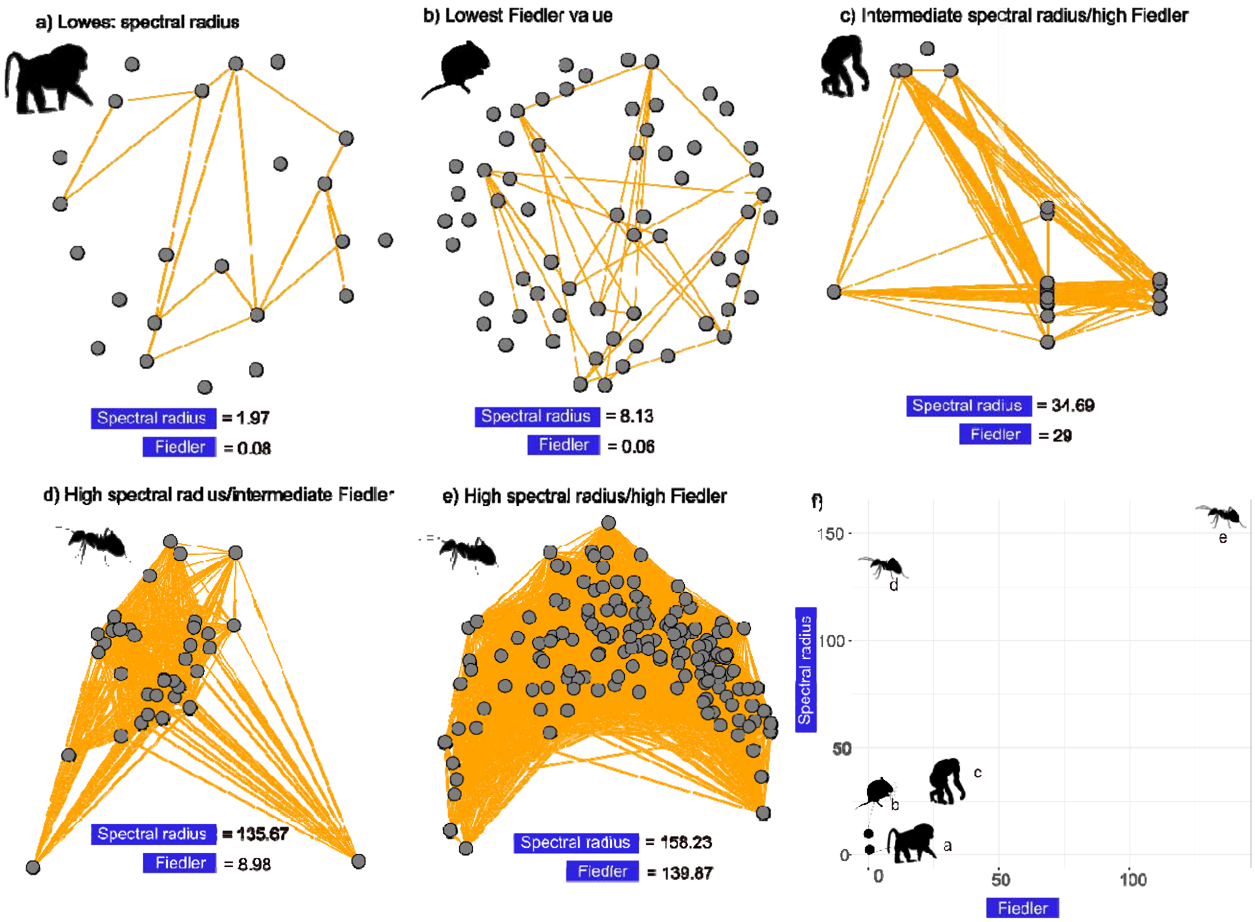

Fig.1: Examples of networks analysed in this study with a) the lowest spectral radius (baboons *Papio cynocephalus* contact network), b) the lowest Fiedler value (voles *Microtus agrestis* trap sharing network), c) intermediate spectral radius values but high Fiedler value (Chimpanzee *Pan troglodytes* contact network), d) high spectral radius/intermediate Fiedler value (*Camponotus fellah* colony contact network) and e) high values of both measures (another *C. fellah* colony contact network). The mean values across all networks were 34.80 and 7.31 for the spectral radius and Fiedler value respectively. f) summary of values across networks (a-e). Silhouettes were sourced from phylopic (http://phylopic.org/). Note that disconnected nodes were not included in the analysis.

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.28.501936; this version posted August 1, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

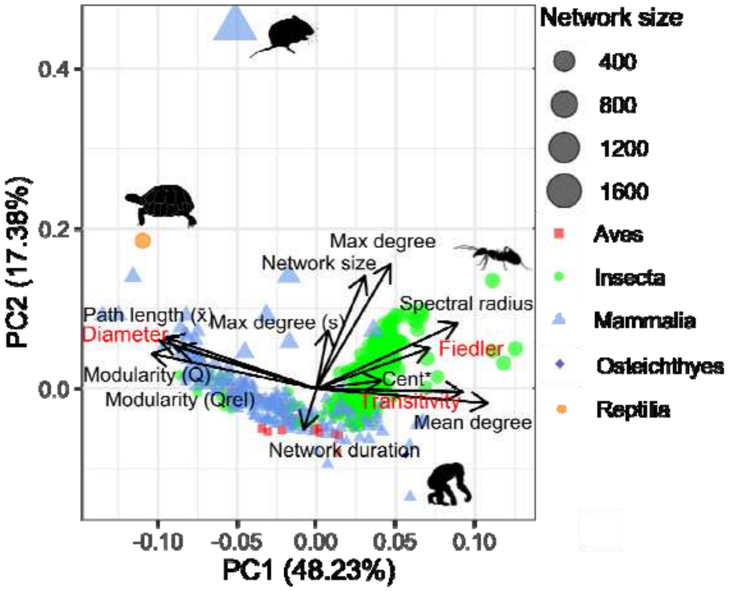

Fig. 2: Principal components analysis (PCA) biplot showing that network structure largely clusters by taxonomic class. Points are coloured by taxa. Points closer together in Euclidean space have networks more similar in structure. Points are scaled by network size. The length and direction of vectors (black arrows) shows how each variable relates to each principal component with larger vectors having higher loadings on that axis. The PCA was constructed just using continuous network characteristics. Percentages next to PC scores indicate how much variability in the data is accounted for by each axis. Cent*: Centralization. See Table S1 for axis loadings and Fig. S1 for the species-level clustering. See Tables S2 & S3 for variable definitions. Silhouettes for some of the outlying networks were sourced from phylopic (http://phylopic.org/). s = scaled. Cent = Centralization.

Spectral properties predict pathogen spread across epidemic scenarios

We found that network characteristics alone could predict pathogen transmission dynamics remarkably well (Figs. 3 & Fig S3). We constructed models in *MrIML* to predict the maximum proportion of nodes infected in the network over 100 time steps (hereafter 'proportion infected'). With these models we could predict the proportion infected in a network using both spectral measures and species identity alone (Fig. 3a). Network size, relative modularity and centralization, for example, were less important in predicting proportion infected across all SIR model parameter combinations tested (Fig. 2a). Nonlinear relationships were likely important for prediction of proportion infected, as random forests

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.28.501936; this version posted August 1, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

(RF) had the highest predictive performance overall (Table S4) and substantially outperformed linear regression in the *MrIML* framework (root mean square error (RMSE) 0.13 vs 0.03). Variable importance and predictor conditional effects were consistent between the machine learning algorithms, so we subsequently analysed the best performing RF model. Across all SIR parameter combinations, we found a nonlinear relationship between proportion infected and spectral radius, with the average prediction of proportion infected increasing by ~30% across the range of spectral radius values (holding all other variables constant in the model, Fig. 3b). In contrast we found a more modest effect of the Fiedler value, with the proportion of infected only increasing on average ~3% across the observed range of values for all SIR parameters (Fig 3c). We did find a sharp increase in the proportion infected in networks when the Fiedler value was less than about 15 (Fig. 3c). However, there was variation in the relationship between the proportion infected and these spectral values across transmission (β) and recovery probabilities (γ, Figs. 3d-e). For example, when the probability of transmission was relatively high ($\beta = 0.2$) and recovery low ($\gamma = 0.04$) the proportion infected across networks was ~80% and spectral radius had a relatively minor effect (Fig. 3d). A network's spectral radius had a stronger effect when the probability of recovery was higher ($\gamma = 0.4$) across all values of β. The increase in proportion infected when the Fiedler value was low (< 15) was not apparent when spread was slower and chances of recovery higher (e.g., $\beta = 0.025$ or 0.01, $\gamma = 0.4$; Fig 3e). The spectral radius and Fiedler value patterns overall were similar, with larger values reducing the time-to-peak prevalence (hereafter 'time to peak', Fig. S3). However, modularity played a greater role in our time to peak models, with the time to peak being longer for more modular networks above a Q_{rel} threshold of ~ 0.75 (Fig. S4).

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.28.501936; this version posted August 1, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

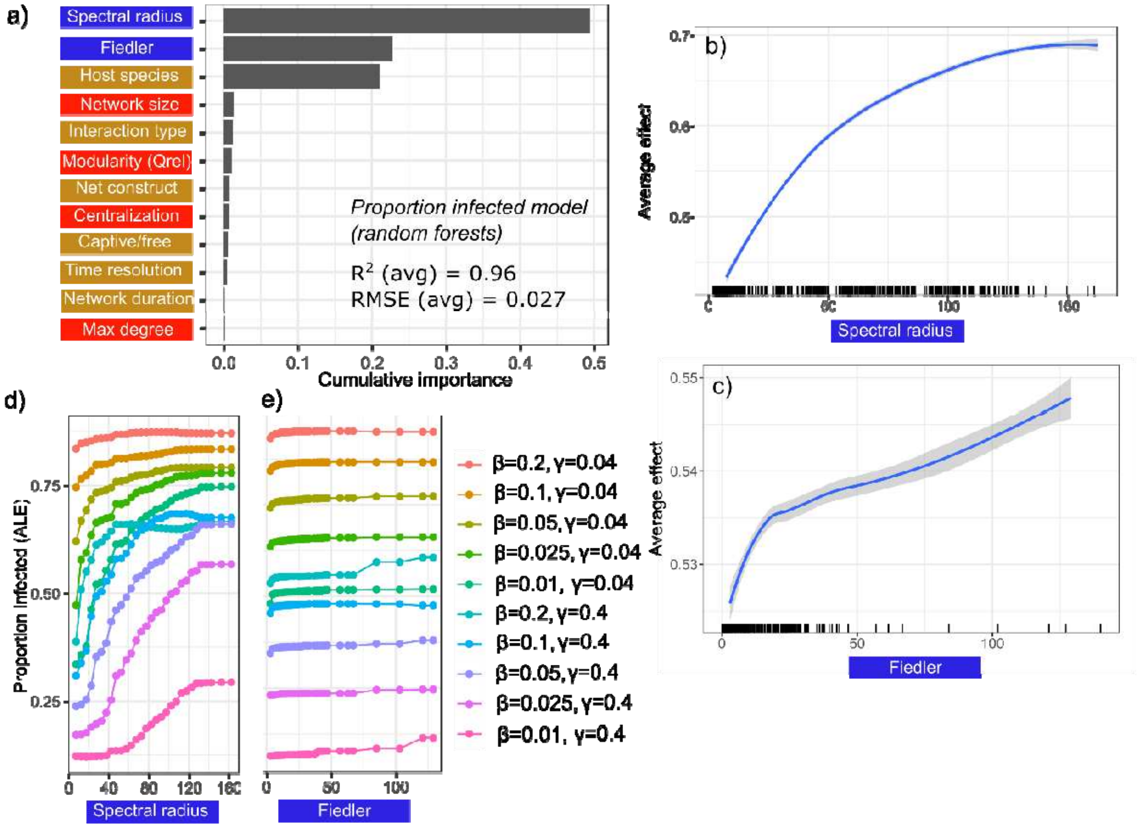

Fig. 3: Plots showing the predictive performance, variable importance and the functional form of relationships for our best-performing *MrIML* proportion infected model. See Table S4 for model performance estimates across algorithms. The colour of the labels indicates what type of predictor it is (blue = spectral, red = non-spectral network structural variables, gold = network metadata, see Tables S2 & S3). a) Spectral radius and the Fiedler value (followed by species) are the most important predictors of proportion of individuals infected across all simulations (importance threshold >0.1) and overall model performance was high (average R^2 = 0.96 and root mean square error (RMSE) = 0.027). b-c) Average predictive surface showing the relationship between spectral properties and proportion infected across all epidemic values (95% confidence intervals in grey). Rug plot on the x axis of the panels on the right shows the distribution of each characteristic across empirical networks. d-e) The accumulated local effects (ALE) plot revealed that the strongly non-linear relationships between both spectral properties and proportion infected were mediated by transmission and recovery probabilities. We chose these SIR parameter values (β = transmission probability, γ = recovery probability) to ensure major outbreaks occurred on the empirical networks. Net construct = Network construction method.

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.28.501936; this version posted August 1, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Simplifying our models with global surrogates

When we further interrogated our moderate (β = 0.05) transmission models, we found that the spectral radius and Fiedler value overall also played a dominant role in our predictions of spread. To quantify the putative mechanisms that underlie our model predictions – 'to decloak the black box' − and gain insight into possible interactions between predictors, we constructed surrogate decision trees as a proxy for our more complex RF model. We trained our surrogate decision tree on the predictions of the RF model rather than the network observations directly. In each case, the surrogate decision tree approximated the predictions of our models (thousands of decision trees) remarkably well (Global $R^2 > 0.95$, see (Molnar 2018) for details). The spectral radius and, to a lesser extent, the Fiedler value and modularity values dominated surrogate trees for all SIR parameter sets (Fig. 5, Figs. S5 & S6). For example, for networks with a Fiedler value ≥ 0.86 and a spectral radius ≥ 20 (as was the case for 51% of our networks, Fig. 4b) the estimated maximum proportion of the network infected was 0.92 (Fig. 4b). The duration over which the data was collected also was included in the surrogate model, with networks collected over > 6.5 days having higher estimates of proportion infected (Fig. 4b).

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.28.501936; this version posted August 1, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

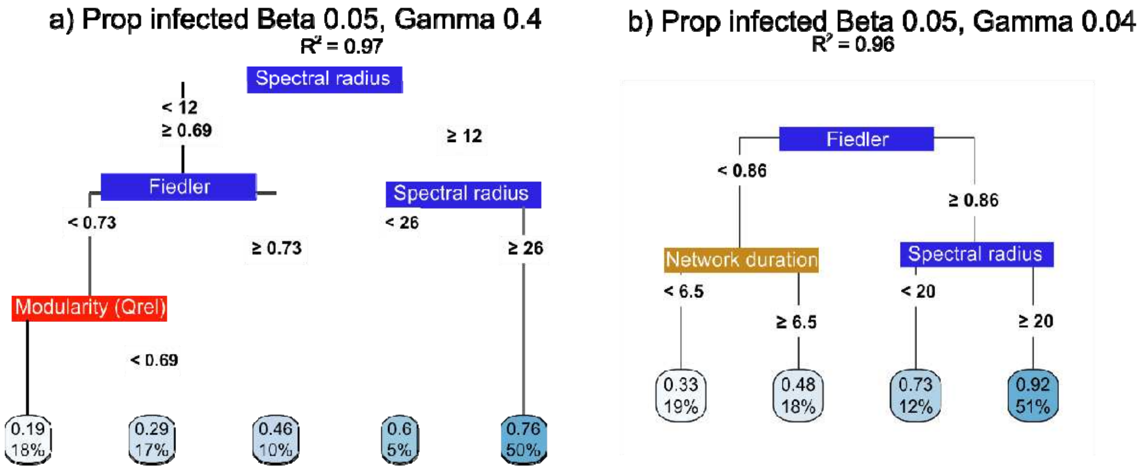

Fig. 4 Global surrogate decision trees for our moderate transmission (β = 0.05) proportion infected with a) high and b) low recovery probability (γ = 0.4 and 0.04 respectively). Threshold values of each variable are included in each tree. The boxes at the tips of the trees indicate the estimates of average peak time or proportion of the network infected across simulations (top value) and percentage of networks in our dataset to be assigned to this tip. For example, 50% of our empirical networks had spectral radius values ≥ 26 and for these networks we found on average, a maximum of 0.76 of the network infected after 100 time steps. Tip boxes are coloured light to dark blue based on network vulnerability to pathogen spread (e.g., longer time to peak = light blue). Global fit = R^2 for how well the surrogate model replicates the predictions of the trained model. See Figs. S5 for the complete list of global surrogate models and Fig. S6 for 'time to peak' surrogates. Colour of the label indicates what type of predictor it is (blue = spectral, red = non-spectral structural variables, gold = network metadata, see Tables S2 & S3).

Do our structural estimates generalize to more complex spread scenarios?

To further validate our predictions, we examined how our models predicted *M. bov*is spread across badger networks with empirical estimates using Shapley values (Shapley, 1951). Shapley values are a game-theoretic approach to explore the relative contribution of each predictor on individual networks (see *Methods*). While *M. bovis* in badgers often has a prolonged latent period and individuals do not typically recover, generally *M. bovis* is a slow-spreading infection, with an R_0 of between 1.1 and 1.3 (Delahay *et al.* 2013). Thus, we interrogated our most similar model (β = 0.05, γ = 0.04, $R_0 = 1.25$). Our model predicted the proportion of infected badgers in the network to be 0.45, which was much lower than the

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.28.501936; this version posted August 1, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

average proportion infected across all networks included in our study (0.71, Fig. 5a). This difference was largely driven by the badger network's low Fiedler value (0.096, much lower than the mean of 7.31 across all networks) and, to a lesser degree, by the small spectral radius (8.10 compared to a mean of 34.8 across all networks, Fig. 5a). This is comparable to contemporaneous estimates of *M. bovis* prevalence in this population, e.g., 41% of badgers tested in the network study tested positive (Weber *et al.* 2013).

We further validated our approach using two Tasmanian devil contact networks (calibrated to reflect potential DFTD transmission) not included in our training data (Fig. 5) and compared to model estimate of spread to empirical observations in similar populations. Based on our model that most closely mirrored devil facial tumour disease DFTD dynamics ($\beta = 0.2$, $\gamma = 0.04$, $R_0 = 5$, see Hamede *et al.* (2012)) we estimated the proportion infected to be 0.85-0.88 for mating and non-mating seasons respectively. Inputting the devil networks' Fiedler value and spectral radius into the corresponding global surrogate model provides an estimate of 0.89 of individuals in the network infected (Fig. 5b). The spectral values were the most important predictors in this model (Fig. 5c). Even though our simulations were not formulated to model DFTD (e.g., devils rarely recover from DFTD), our machine-learning estimates closely predicted the empirical findings for this disease. In comparable populations across the island where the disease was monitored before the onset of the disease, maximum prevalence estimates ranged from 0.7-1.0 in for sexually matured devils (≥ 2 y.o.) ~100 weeks after disease arrival (McCallum *et al.* 2009). Our predictions of proportion infected were not particularly sensitive to transmissibility estimates as in our model. For example, with a 50% reduction in the probability of transmission ($\beta = 0.1$) our estimate of proportion infected was still similar to empirical estimates (0.83, Fig. S5a). Taken together, our findings

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.28.501936; this version posted August 1, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

show how the spectral values of contact networks offer a valuable and informative “shorthand” for how vulnerable different animal networks are to outbreaks.

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.28.501936; this version posted August 1, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

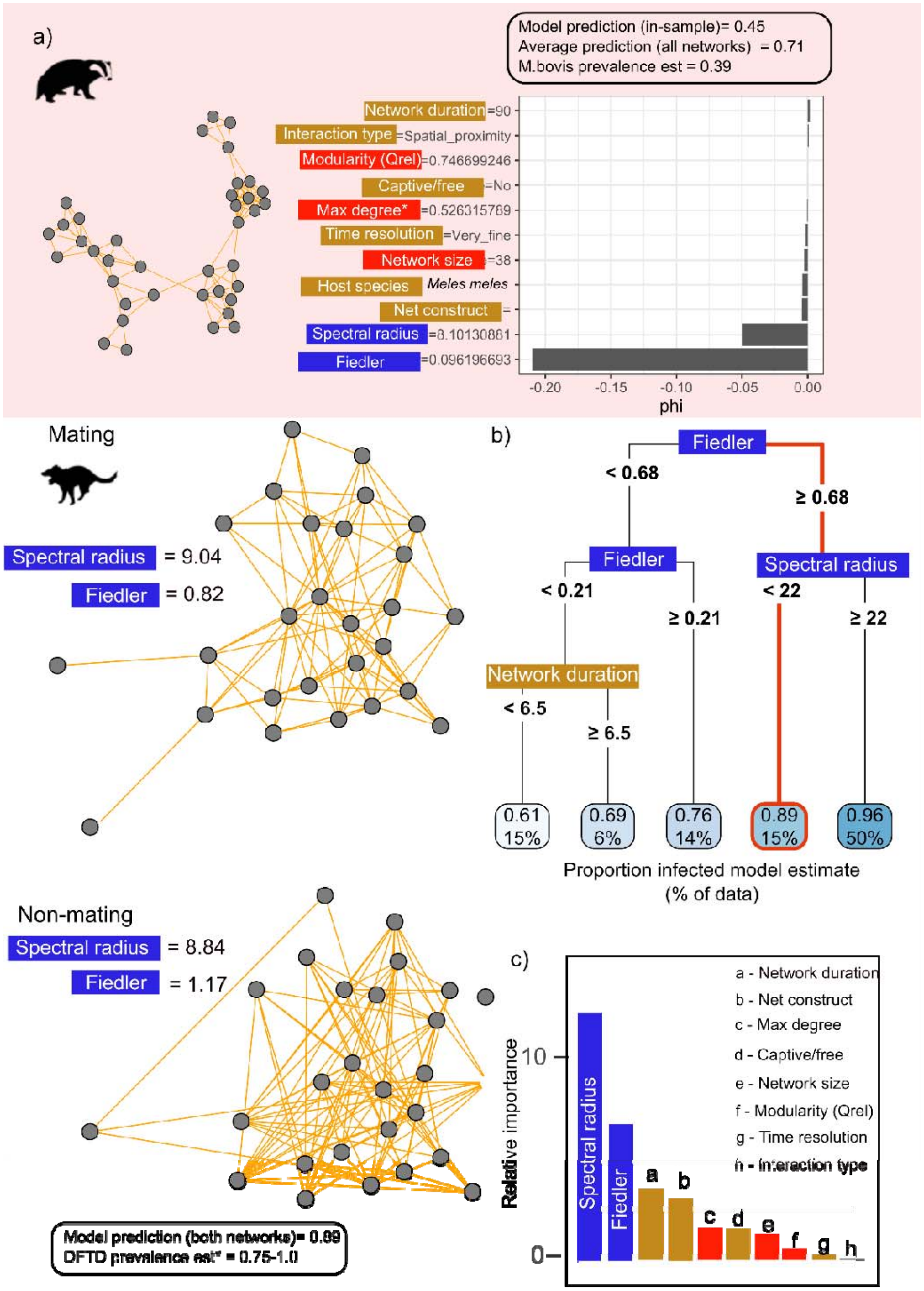

Fig. 5: The spectral radius and the Fiedler underpinned our in-sample prediction of the proportion infected estimates in our a) badger and b/c) out-of-sample Tasmanian devil contact networks. a) Shapley values (φ) that quantify how each variable shaped simulated proportion infected ($\beta = 0.05$, $\gamma = 0.04$) in an empirical badger network. Negative Shapley values indicate that the variable reduced the proportion infection relative to other variables included in the model. See Fig S7 for other Shapley value analyses of other contrasting networks. b) Surrogate decision tree for the model that best approximated Tasmanian devil facial tumour disease (DFTD, $\beta = 0.2$, $\gamma = 0.04$). Red lines indicate the branches of the tree

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.28.501936; this version posted August 1, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

corresponding to the spectral values from the left panels. The red outlined box is the estimated proportion infected for both networks. c) Corresponding variable importance plot showing the spectral radius and Fiedler value followed by data duration were the most important predictors in the model. Colour of the labels indicates what type of predictor it is (blue = spectral, red = non-spectral structural variables, gold = network metadata, see Tables S2/S3). Panels on the left are the corresponding networks. Net construct = Network construction method. *: for sexually mature individuals in comparable populations over similar time scales to the simulations (McCallum *et al.* 2009).

Discussion

Here, we show that the spectral radius and Fiedler value of a network can be a remarkably strong predictor for population vulnerability to diverse epidemics varying in key epidemiological parameters. We demonstrate how a powerful machine learning and simulation approach can effectively predict pathogen outbreak dynamics on a large collection of empirical animal contact networks. We not only demonstrate the high predictive power of a network's spectral properties but also show that our predictions can be a useful tool for estimating spread in systems with complex disease dynamics. Our findings offer insights into how nuances in social organisation translate into differences in pathogen spread across the animal kingdom. Furthermore, our global surrogate models provide animal health practitioners with an intuitive framework to gain rapid insights into the vulnerability of populations to the spread of emerging infectious diseases.

Across real-world contact networks, we found that the networks' spectral properties (Fiedler distance and spectral radius) were powerful proxies for pathogen spread. The strong relationship between spectral radius and epidemic threshold has been demonstrated for theoretical networks (Prakash *et al.* 2010) and has been used to assess vulnerability of cattle movement networks to spread of bovine brucellosis (Darbon *et al.* 2018). We expand these findings to show that the spectral radius is the most important predictor in our models of epidemic behaviour across diverse animal social systems. While we examined only SIR

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.28.501936; this version posted August 1, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

propagation through our networks, theoretical results suggest that our findings will extend to other propagation mechanics such as SIS, (susceptible-infected-susceptible) and SEIR (susceptible, exposed, infected, recovered) (Prakash *et al.* 2010). Given that both the badger *M. bovis* and DFTD systems have more complex propagation mechanics compared to SIR, our models could still predict disease dynamics of both disease systems reasonably well. We that note that for DFTD, disease simulation models that assume homogeneous mixing of hosts provide similar estimates of disease dynamics to network-based simulations (Hamede *et al.* 2012). However, Hamede *et al.* (2012) found the outcome of simulated DFTD epidemics sensitive to estimates of latent period and transmissibility parameters, whereas our network structure approach provided realistic estimates of prevalence with minimal reliance on parameter values.

For some networks and epidemiological parameters, spectral radius alone was not sufficient to predict spread, and the Fiedler value and modularity still played an important role. The Fiedler value and spectral radius of the networks were correlated, but below our $\rho = 0.7$ threshold (Fig. S2). One potential reason for this is that the Fiedler value seems to be less sensitive to nodes with high connectivity compared to the spectral radius (Fig. 1); however, the mathematical relationship between these two algebraic measures of connectivity is poorly understood (Tang & Priebe 2016). Combined, our global surrogate models and accumulated effects plots pointed to networks such as the devil networks with spectral radii > ~8 and Fiedler values > 1 being more vulnerable to pathogen spread (the effect of the Fiedler value on spread was much weaker overall). The spectral properties were dominant for the fast-spreading pathogen models (e.g., example system), whereas network size and modularity played a more important role in our models for more slowly spreading pathogens (e.g., Figs. S5 & S6).

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.28.501936; this version posted August 1, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

When modular structure played a role in disease spread in our study, we detected similar patterns to those found by Sah *et al.* (2017). As in Sah *et al.* (2017), we found that epidemic progression was only slowed in highly modular networks ($Q_{rel} > \sim 0.7$) when the probability of transmission between nodes was low ($\beta > 0.025$). Such subdivided networks were rare in our data and are commonly associated with high fragmentation (small groups or sub-groups) and high subgroup cohesion (Sah *et al.* 2017). The reduced importance of modularity relative to spectral radius is due to within-group connections being crucial for epidemic outcomes in many contexts (Sah *et al.* 2017). Spectral values may have higher predictive performance, as they summarize connectivity across the networks including between- and within-group connections. Interpreting how modularity alone impacted epidemic outcomes was difficult on these empirical networks, as all modularity measures were strongly correlated with mean degree, diameter and transitivity (Fig. 2, Fig. S2). The extent of these correlations can vary wildly based on other aspects of network structure and they all have interacting effects on disease dynamics (Zhang & Zhang 2009; Ames *et al.* 2011). However, the spectral radius captures epidemiologically important aspects of network structure on its own without having to untangle whether different aspects of network structure are correlated.

More broadly, our study provides a framework for how interpretable machine learning can predict spread across networks for a wide variety of epidemic parameters. While our RF *MrIML* model had much higher predictive performance compared to the corresponding linear models, further investigation of these models provided critical insight into how network structure impacted pathogen spread. This framework could identify general trends of disease vulnerability, specific thresholds for pathogens with certain characteristics, as well as the drivers of spread for individual networks.

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.28.501936; this version posted August 1, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

To help practitioners apply our model to different host-pathogen systems, we developed an R-Shiny app (https://spreadpredictr.shinyapps.io/spreadpredictr/). Our web app allows users to make predictions of spread for diverse transmission and recovery probabilities on a contact network of interest without the need for simulation. Even when the underlying mechanism of spread was mis-specified, as with our case studies, our model could provide reasonable estimates of the proportion of the population infected that align closely with empirical data. While currently limited to pathogens with SIR transmission dynamics, future versions of the app will include, for example, SI and SEIR mechanics. We stress that for practitioners to make accurate predictions for a particular pathogen, contact definitions and the duration of data should be calibrated or multiple thresholds for what constitutes a transmission contact assessed (see Craft 2015). For example, for the giraffe network we included edges that represented individuals seen once together over a period of a year, and predictions of pathogen spread on this network would likely be inflated for pathogens requiring more sustained contact (VanderWaal *et al*. 2014). Nonetheless, this study shows the utility of linking network simulation and interpretable machine learning approaches to tease apart the drivers of spread across empirical wildlife networks

As this is a broad, comparative study of simulated pathogen spread on 603 empirical networks across taxonomic groups, we made important simplifying assumptions. For example, as there were large differences in how the empirical network edges were weighted across taxa (e.g., some networks were weighted by contact duration and others by contact frequency) our approach treated all contacts as equal in unweighted networks, as is done in similar studies (Ames *et al.* 2011; Sah *et al.* 2017). We also simulated spread across static networks, making the assumptions (i) that aggregated networks are representative or social

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.28.501936; this version posted August 1, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

patterns at epidemiologically-relevant timescales and (ii) that network change happens more slowly than pathogen spread. Including predictions of spread that account for the dynamic nature of contact structure and pathogen-mediated changes in behaviour is an important future extension of this work. However, applying dynamic network models such as temporal exponential random graph models (Krivitsky & Handcock 2014) to estimate spread is computationally demanding and challenging in a comparative setting due to idiosyncrasies in the model-fitting process. While of high predictive value, our models did not capture all aspects of uncertainty. For example, we assumed each network was fully described, with no missing nodes or edges, which is almost always not the case for wildlife studies. How sensitive spectral properties are to missing data is an open question. However, promisingly, removing edges from ecological networks with high Fiedler values does not appear to strongly impact the stability of the network (Kumar *et al.* 2019).

Another limitation of this study is that our models did not account for uncertainty in predictions. Currently, more probabilistic models such as BART (Bayesian Additive Regression Trees) (Carlson 2020) are not available in the *MrIML* framework, but future extensions may allow for methods such as BART to be incorporated (Fountain-Jones *et al.* 2021). However, one advantage of our approach is that for the RF model (proportion infected), host species (and the other categorical variables, see Table S3) could be added as a categorical predictor rather than hot-encoded set of 43 predictors (one binary predictor for each species (-1)). This simplified interpretations about how host species affect pathogen spread differently, while accounting for nonindependence of intra-species networks (e.g., networks for host species A from different populations of that species or from different timepoints) (Sah *et al.* 2019). A large proportion of the networks (~150) came from one taxon (*C. fellah*); removing this one taxon did not qualitatively change our findings. While this

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.28.501936; this version posted August 1, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

study demonstrates the power of repositories such as the ASNR, there are large biases in the taxa covered that must be accounted for in model structure. Starting to fill in these taxonomic gaps in a systematic way will increase the utility of comparative approaches such as ours and make them generalizable across taxa and populations.

This paper provides a significant step towards a spectral understanding of pathogen spread in animal networks. In particular, we show that the spectral radius of an animal network is a powerful predictor of spread for diverse hosts and pathogens that can be a valuable shortcut for stakeholders to understand the vulnerability of animal networks to disease. We also demonstrate how multivariate interpretable machine learning models can provide novel insights into spread across scales. Moreover, this study identified the key axes of network structural variation across the animal kingdom that can inform future comparative network research. As rapid advances in location-based tracking and bio-logging (Katzner & Arlettaz 2020) make network data more readily available to wildlife managers, approaches like this one will be of increasing value.

Methods

Networks

We downloaded all animal contact networks from the ASNR on 12th January 2022 (Sah *et al.* 2019) and combined these with other comparable published animal contact networks (Rushmore *et al.* 2013; Weber *et al.* 2013; VanderWaal *et al.* 2014). We binarized each network, extracted the largest connected component, and excluded networks with fewer than 10 individuals. This left us with 603 networks from 43 species.

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.28.501936; this version posted August 1, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

From each network we calculated a variety of network structure variables using the R package *igraph* (Csárdi & Nepusz 2006) (see Table S2). As these networks were constructed using a wide variety of techniques, we also extracted metadata from the ASNR or the publication associated with the network (Table S3). These variables were also added to the models. We used Principal Components Analysis (PCA) biplots to examine the drivers of variation in network structure and visualise how networks clustered by taxonomic class. We removed networks with missing metadata (8 networks) and screened for correlations between variables. As many of the machine learning variables are less sensitive to collinearity (Fountain-Jones *et al.* 2019) we used a pairwise correlation threshold of 0.7 and removed variables from the pair with the highest overall correlation (Table S2).

Simulations

To simulate the spread of infection on each network we used our R package "EpicR" (Epidemics by computers in R; available on GitHub at https://github.com/mcharleston/epicr).

The simulations use a standard discretisation of the SIR model, in which time proceeds in "ticks," for example representing days. Initially one individual was chosen at uniform random to be infected (I) and all others were susceptible (S). At each time step, one of two changes of state can happen to each individual (represented by a node), depending on its current state. An 'S' individual will become infected (I) with a probability $(1 - (1 - \beta)^k)$, where k is the number of currently infected neighbours it has, or otherwise stay as S; an 'I' individual will recover (R) with probability γ or remain as I. Recovered (R) individuals stay as R.

In classical deterministic SIR models as a set of differential equations, β and γ are instantaneous rates; here, they are probabilities per time step, so at a coarse level, they are comparable.

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.28.501936; this version posted August 1, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

On each network, we performed 1000 simulations using different combinations of transmission (β = 0.01, 0.025, 0.05, 0.1, 0.2) and recovery probabilities (γ = 0.04, 0.4). We chose these values to broadly reflect a range of scenarios from high to low transmissibility and slow to fast recovery (Leung 2021) and ensure large outbreaks (>10% on individuals infected, see Fig S8 for the analysis with a wider variety of recovery rates) (Sah *et al.* 2017). For each simulation we recorded two complementary epidemic measures to capture disease burden and speed of spread: a) the maximum prevalence reached, or the maximum proportion of individuals infected in the network after 100 time steps and b) time to outbreak peak (i.e., which time step had the maximum number of infections). We chose 100 time steps to ensure that the epidemic ended and there were no remaining infected nodes. One randomly chosen individual was infected at the beginning of the simulation. The average maximum proportion infected and time to outbreak across all simulations for each parameter combination were used as the response variables in the machine learning models,

Machine learning pipeline

We used a recently developed multi-response interpretable machine learning approach *(Mr IML*, Fountain-Jones *et al.* 2020) to predict outbreak characteristics using network structure variables. Our *MrIML* approach had the advantage of allowing us to rapidly construct and compare models across a variety of machine-learning algorithms for each of our response variables as well as assess generalized predictive surfaces across epidemic parameters.

To test the robustness of our results, we compared the performance of four different underlying supervised regression algorithms in our *MrIML* models. We compared linear models (LMs), support vector machines (SVMs), random forests (RF) and gradient boosted

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.28.501936; this version posted August 1, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

models (GBMs) as they operate in markedly different ways that can affect predictive performance (Fountain-Jones *et al.* 2019; Machado *et al.* 2019). Categorical predictors such as ‘species’ were hot-encoded for some models as needed (see Table S4). As both types of responses in our models were continuous, we compared the performance of each algorithm using the average R^2 and root mean squared error (RMSE) across all responses (hereafter, the ‘global model’). As we included models that were not fit using sums of squares, our R^2 estimate depended on the squared correlation between the observed and predicted values (Kvålseth 1985). As ants (Insecta: Formicinae) were over-represented, we compared model performance and interpretation with and without these networks. To calculate each performance metric, we used 10-fold cross validation to prevent overfitting each model. We tuned hyperparameters for each model (where appropriate) using 100 different hyper-parameter combinations (a 10×10 grid search) and selected the combination with the lowest RMSE. The underlying algorithm with the highest predictive performance was interrogated further.

We interpreted this final model using a variety of model-agnostic techniques within the *MrIML* framework. We assessed overall and model-specific variable importance using a variance-based method (Greenwell *et al.* 2018). We quantified how each variables alters epidemic outcomes using accumulated local effects (ALEs) (Apley & Zhu 2016). In brief, ALEs isolate the effect of each network characteristic on epidemic outcomes using a sliding window approach calculating the average change in prediction across the values range (while holding all other variables constant) (Molnar 2018). ALEs are less sensitive to correlations and straightforward to interpret as points on the ALE curve are the difference from the mean prediction (Apley & Zhu 2016; Molnar 2018; Fountain-Jones *et al.* 2021).

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.28.501936; this version posted August 1, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

To further examine the predictive performance of our black-box models (SVM, RF and GBM) we calculated a global surrogate decision tree (hereafter 'global surrogate') to approximate the predictions of our more complex trained models. Global surrogates are generated by training a simpler decision tree to the *predictions* (instead of observations) of the more complex 'black box' models using the network structure data. How well the surrogate model performed compared to the complex model is then estimated using R^2. See Molnar (2018) for details.

Lastly, we gained more insight into model behaviour and how network structure impacted epidemic outcomes on individual networks, including by calculating Shapley values (Štrumbelj & Kononenko 2014). Shapley values use a game theoretic approach to play off variables in the model with each other based on their contribution to the prediction (Shapley 1953). For example, negative Shapley values indicate that the observed value 'contributed to the prediction' by reducing the proportion infected or time to peak in an outbreak for a particular network. See Molnar (2018) for a more detailed description and (Fountain-Jones *et al.* 2019; Worsley-Tonks *et al.* 2020) for how they can be interpreted in epidemiological settings.

We validated our results using networks with well-documented disease dynamics. The European badger network was included in our training data, and we selected the propagation model with a slow recovery rate ($\gamma = 0.04$) and intermediate transmissibility ($\beta = 0.05$) that provided an equivalent/similar R_0 (1.1-1.3) to *M. bovis* in the studied badger population (Delahay *et al.* 2013). It should be noted here that *M. bovis* infection has SEI(R)(D)

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.28.501936; this version posted August 1, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

dynamics, being frequently latent in badgers for long periods with infection only resolving in some individuals (the most infectious individuals with progressed disease have elevated mortality (Corner *et al.* 2011)). We compared the proportion infected returned by our model to various contemporaneous estimates of *M. bovis* prevalence (Delahay *et al.* 2013; Buzdugan *et al.* 2017) in the long-term study that contact network data were collected in (McDonald *et al.* 2018).

The Tasmanian devil networks were not included in the training data. To compare predictions, we extracted the predict function from the model that was the most similar to estimates of DFTD dynamics based on empirical data ($\beta = 0.2$, $\gamma = 0.04$, $R_0 = 5$) (McCallum *et al.* 2009; Hamede *et al.* 2012). DFTD has SEI(D) dynamics in devil populations, however, accurately estimating the latent period is impossible, as there is (as of May 2022) no diagnostic tool to detect DFTD prior to visual detection of the tumours (Hamede *et al.* 2012). As we wanted to make predictions on a species not included in our dataset, we reran the models excluding the species predictor and the model performance, and results were very similar. See https://github.com/nfj1380/igraphEpi for our complete analytical pipeline.

Acknowledgements

This project was supported by an Australian Research Council Discovery Project Grant (DP190102020). We would also like to thank Prof. Menna Jones and Prof. Sue VandeWoude for their support and comments on this manuscript.

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.28.501936; this version posted August 1, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

References

Ames, G.M., George, D.B., Hampson, C.P., Kanarek, A.R., McBee, C.D., Lockwood, D.R., *et al.* (2011). Using network properties to predict disease dynamics on human contact networks. *Proceedings of the Royal Society B: Biological Sciences*, 278, 3544–3550.

Apley, D.W. & Zhu, J. (2016). Visualizing the effects of predictor variables in black box supervised learning models. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, 82, 1059–1086.

Bansal, S., Grenfell, B.T. & Meyers, L.A. (2007). When individual behaviour matters: homogeneous and network models in epidemiology. *J R Soc Interface*, 4, 879–891.

Bansal, S., Read, J., Pourbohloul, B. & Meyers, L.A. (2010). The dynamic nature of contact networks in infectious disease epidemiology. *J Biol Dyn*, 4, 478–489.

Buzdugan, S.N., Vergne, T., Grosbois, V., Delahay, R.J. & Drewe, J.A. (2017). Inference of the infection status of individuals using longitudinal testing data from cryptic populations: Towards a probabilistic approach to diagnosis. *Sci Rep*, 7, 1111.

Carlson, C.J. (2020). embarcadero: Species distribution modelling with Bayesian additive regression trees in R. *Methods in Ecology and Evolution*, 11, 850–858.

Corner, L.A.L., Murphy, D. & Gormley, E. (2011). *Mycobacterium bovis* Infection in the Eurasian Badger (*Meles meles*): the disease, pathogenesis, epidemiology and control. *Journal of Comparative Pathology*, 144, 1–24.

Craft, M.E. (2015). Infectious disease transmission and contact networks in wildlife and livestock. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 370, 20140107-.

Craft, M.E., Volz, E., Packer, C. & Meyers, L.A. (2011). Disease transmission in territorial populations: the small-world network of Serengeti lions. *Journal of the Royal Society Interface*, 8, 776–786.

Darbon, A., Valdano, E., Poletto, C., Giovannini, A., Savini, L., Candeloro, L., *et al.* (2018). Network-based assessment of the vulnerability of Italian regions to bovine brucellosis. *Preventive Veterinary Medicine*, 158, 25–34.

Delahay, R.J., Walker, N., Smith, G.S., Wilkinson, D., CLIFTON-HADLEY, R.S., Cheeseman, C.L., *et al.* (2013). Long-term temporal trends and estimated transmission rates for Mycobacterium bovis infection in an undisturbed high-density badger (Meles meles) population. *Epidemiology and Infection*, 141, 1445–1456.

Eames, K., Bansal, S., Frost, S. & Riley, S. (2015). Six challenges in measuring contact networks for use in modelling. *Epidemics*, 10, 72–77.

Fountain-Jones, N., Kozakiewicz, C., Forester, B., Landguth, E., Carver, S., Charleston, M., *et al.* (2021). MrIML: Multi-response interpretable machine learning to map genomic landscapes. *Molecular Ecology Resources*, 21, 2766–2781.

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.28.501936; this version posted August 1, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Fountain-Jones, N.M., Machado, G., Carver, S., Packer, C., Recamonde-Mendoza, M. & Craft, M.E. (2019). How to make more from exposure data? An integrated machine learning pipeline to predict pathogen exposure. *Journal of Animal Ecology*, 88.

Greenwell, B.M., Boehmke, B.C. & Mccarthy, A.J. (2018). A simple and effective model-based variable importance measure. *arXiv Preprint arXiv:1805.04755*.

Hamede, R., Bashford, J., Jones, M. & McCallum, H. (2012). Simulating devil facial tumour disease outbreaks across empirically derived contact networks. *Journal of Applied Ecology*, 49, 447–456.

Hamede, R.K., Bashford, J., McCallum, H. & Jones, M. (2009). Contact networks in a wild Tasmanian devil (*Sarcophilus harrisii*) population: using social network analysis to reveal seasonal variability in social behaviour and its implications for transmission of devil facial tumour disease. *Ecology Letters*, 12, 1147–1157.

Katzner, T.E. & Arlettaz, R. (2020). Evaluating contributions of recent tracking-based animal movement ecology to conservation management. *Frontiers in Ecology and Evolution*, 7.

Krivitsky, P.N. & Handcock, M.S. (2014). A Separable Model for Dynamic Networks. *J R Stat Soc Series B Stat Methodol*, 76, 29–46.

Kumar, D., Gupta, J. & Raha, S. (2019). Partitioning a reaction–diffusion ecological network for dynamic stability. *Proc Math Phys Eng Sci*, 475, 20180524.

Kvålseth, T.O. (1985). Cautionary Note about R 2. *The American Statistician*, 39, 279–285.

Lentz, H.H.K., Selhorst, T. & Sokolov, I.M. (2012). Spread of infectious diseases in directed and modular metapopulation networks. *Phys Rev E Stat Nonlin Soft Matter Phys*, 85, 066111.

Machado, G., Vilalta, C., Recamonde-Mendoza, M., Corzo, C., Torremorell, M., Perez, A., *et al.* (2019). Identifying outbreaks of Porcine Epidemic Diarrhea virus through animal movements and spatial neighborhoods. *Scientific Reports*, 9, 457.

McCabe, C.M. & Nunn, C.L. (2018). Effective network size predicted from simulations of pathogen outbreaks through social networks provides a novel measure of structure-standardized group size. *Frontiers in Veterinary Science*, 5.

McCallum, H., Jones, M., Hawkins, C., Hamede, R., Lachish, S., Sinn, D.L., *et al.* (2009). Transmission dynamics of Tasmanian devil facial tumor disease may lead to disease-induced extinction. *Ecology*, 90, 3379–3392.

McDonald, J.L., Robertson, A. & Silk, M.J. (2018). Wildlife disease ecology from the individual to the population: Insights from a long-term study of a naturally infected European badger population. *Journal of Animal Ecology*, 87, 101–112.

Meyers, L.A., Newman, M.E.J. & Pourbohloul, B. (2006). Predicting epidemics on directed contact networks. *J Theor Biol*, 240, 400–418.

Meyers, L.A., Pourbohloul, B., Newman, M.E.J., Skowronski, D.M. & Brunham, R.C. (2005). Network theory and SARS: predicting outbreak diversity. *J Theor Biol*, 232, 71–81.

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.28.501936; this version posted August 1, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Molnar, C. (2018). *Interpretable machine learning*. Retrieved from https://christophm.github.io/interpretable-ml-book/.

Newman, M.E.J. (2002). The spread of epidemic disease on networks. *Physical Review E*, 66.

Newman, M.E.J. (2006). Community structure in social and biological networks. *Proceedings of the National Academy of Sciences of the United States of America*, 99, 7821–6.

Porter, M.A. (2020). Nonlinearity + Networks: A 2020 Vision. *Emerging Frontiers in Nonlinear Science*, 32, 131–159.

Prakash, B.A., Chakrabarti, D., Faloutsos, M., Valler, N. & Faloutsos, C. (2010). Got the Flu (or Mumps)? Check the Eigenvalue! *arXiv:1004.0060 [cond-mat, physics:physics, q-bio]*.

Rushmore, J., Caillaud, D., Matamba, L., Stumpf, R.M., Borgatti, S.P. & Altizer, S. (2013). Social network analysis of wild chimpanzees provides insights for predicting infectious disease risk. *Journal of Animal Ecology*, 82, 976–986.

Sah, P., Leu, S.T., Cross, P.C., Hudson, P.J. & Bansal, S. (2017). Unraveling the disease consequences and mechanisms of modular structure in animal social networks. *PNAS*, 114, 4165–4170.

Sah, P., Mann, J. & Bansal, S. (2018). Disease implications of animal social network structure: A synthesis across social systems. *Journal of Animal Ecology*, 87, 546–558.

Sah, P., Méndez, J.D. & Bansal, S. (2019). A multi-species repository of social networks. *Sci Data*, 6, 44.

Sah, P., Otterstatter, M., Leu, S.T., Leviyang, S. & Bansal, S. (2021). Revealing mechanisms of infectious disease spread through empirical contact networks. *PLOS Computational Biology*, 17, e1009604.

Salathé, M. & Jones, J.H. (2010). Dynamics and Control of Diseases in Networks with Community Structure. *PLOS Computational Biology*, 6, e1000736.

Shapley, L.S. (1953). Stochastic games. *Proceedings of the National Academy of Sciences of the United States of America*, 39, 1095–1100.

Shu, P., Wang, W., Tang, M., Zhao, P. & Zhang, Y.-C. (2016). Recovery rate affects the effective epidemic threshold with synchronous updating. *Chaos*, 26, 063108.

Silk, M.J., Croft, D.P., Delahay, R.J., Hodgson, D.J., Boots, M., Weber, N., *et al.* (2017). Using Social Network Measures in Wildlife Disease Ecology, Epidemiology, and Management. *Bioscience*, 67, 245–257.

Štrumbelj, E. & Kononenko, I. (2014). Explaining prediction models and individual predictions with feature contributions. *Knowledge and Information Systems*, 41, 647–665.

Tang, M. & Priebe, C.E. (2016). Limit theorems for eigenvectors of the normalized Laplacian for random graphs. *arXiv:1607.08601 [stat]*.

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.28.501936; this version posted August 1, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

VanderWaal, K.L., Atwill, E.R., Isbell, L.A. & McCowan, B. (2014). Linking social and pathogen transmission networks using microbial genetics in giraffe (*Giraffa camelopardalis*). *The Journal of Animal Ecology*, 83, 406–14.

Weber, N., Carter, S.P., Dall, S.R.X., Delahay, R.J., McDonald, J.L., Bearhop, S., *et al.* (2013). Badger social networks correlate with tuberculosis infection. *Current Biology*, 23, R915–R916.

White, L.A., Forester, J.D. & Craft, M.E. (2017). Using contact networks to explore mechanisms of parasite transmission in wildlife. *Biological Reviews*, 92, 389–409.

Worsley-Tonks, K.E.L., Escobar, L.E., Biek, R., Castaneda-Guzman, M., Craft, M.E., Streicker, D.G., *et al.* (2020). Using host traits to predict reservoir host species of rabies virus. *PLOS Neglected Tropical Diseases*, 14, e0008940.

Zhang, Z. & Zhang, J. (2009). A big world inside small-world networks. *PLoS One*, 4, e5686.