Identifying the dynamic contact network of infectious disease spread

Pratha Sah[1] **and Shweta Bansal**[1*]

***For correspondence:**
shweta.bansal@georgetown.edu
(SB)

[1]Department of Biology, Georgetown University, Washington, DC

Abstract

Pathogen propagation is a fundamental process that takes place through host contact networks. While it has been possible to characterize contact networks of several human infectious diseases (e.g., sexual contacts for HIV, physical proximity for measles), logistical difficulties and the high costs of data collection makes network modeling of infectious diseases in animal populations particularly difficult. In addition, limited knowledge about how pathogens transmit precludes an accurate definition of network edges required for constructing networks of disease transmission. We developed a new tool, INoDS (Identifying Networks of infectious Disease Spread), that utilizes Bayesian inference to provide evidence towards the underlying contact network of an infectious disease spread. We show that the tool accurately identifies the underlying contact network even when the networks are partially sampled and information on disease spread is incomplete. We next demonstrate the applicability of the tool in two real animal populations: bumble bees and Australian sleepy lizards. The performance of INoDs in synthetic and complex empirical systems highlights its role as an alternative approach to laboratory transmission experiments, in providing epidemiological insights into novel and less known host-pathogen systems, and overcoming common data-collection constraints.

Introduction

Host contacts, whether direct or indirect, play a fundamental role in the spread of infectious disease through host populations (***Sah et al., 2017***). 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 (***Bansal et al., 2007***). Therefore, in recent years, the network approach to modeling infection spread has gained popularity because it explicitly incorporates 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) data on 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 contact between two individuals and the rate of recovery of infected individuals. The use of modern technology in recent years, including RFID tags, 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 (***Welch et al., 2011***; ***Godfrey, 2013***). More importantly, limited knowledge about the pathogen makes it challenging to identify the mode

of infection transmission, and a disease causing contact between individuals. Laboratory techniques of unraveling transmission mechanisms usually take years to resolve. Defining accurate contact networks underlying infection transmission in human infectious disease has been far from trivial (***Bansal et al., 2007***). 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 (***Craft and Caillaud, 2011***).

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 (***Godfrey et al., 2009***, ***2010***; ***MacIntosh et al., 2012***). However, this approach can be misleading as the position of individuals in social networks depends on dominance hierarchy, age, sex, reproductive status, and can be very different from their position in a contact network that is relevant to infection transmission. 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 (***Danon et al., 2011***; ***Hamede et al., 2009***; ***Fenner et al., 2011***). A recent approach, called the k-test procedure, explores a direct link between infectious disease spread and contact network by comparing the number of infectious contacts of infected cases to that of uninfected cases (***VanderWaal et al., 2016***).

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 (***Aiello et al., 2016***). Furthermore, the relevance of low-weighted edges in a contact network is ambiguous (***Pellis et al., 2014***). 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 evolution of contacts between hosts, however is crucial in determining the order in which contacts occur, which regulates the spread of an infectious disease spread through the host population (***Bansal et al., 2010***). Therefore, a static representation of host interactions often serves as an inferior contact network model because it fails to approximate disease spread through dynamic host networks (***Fefferman and Ng, 2007***; ***Farine, 2017***). 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 the need of 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 Networks of infectious Disease Spread) that utilizes Bayesian inference to identify the underlying contact network of disease transmission in host populations. Our tool can infer contact networks 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. 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

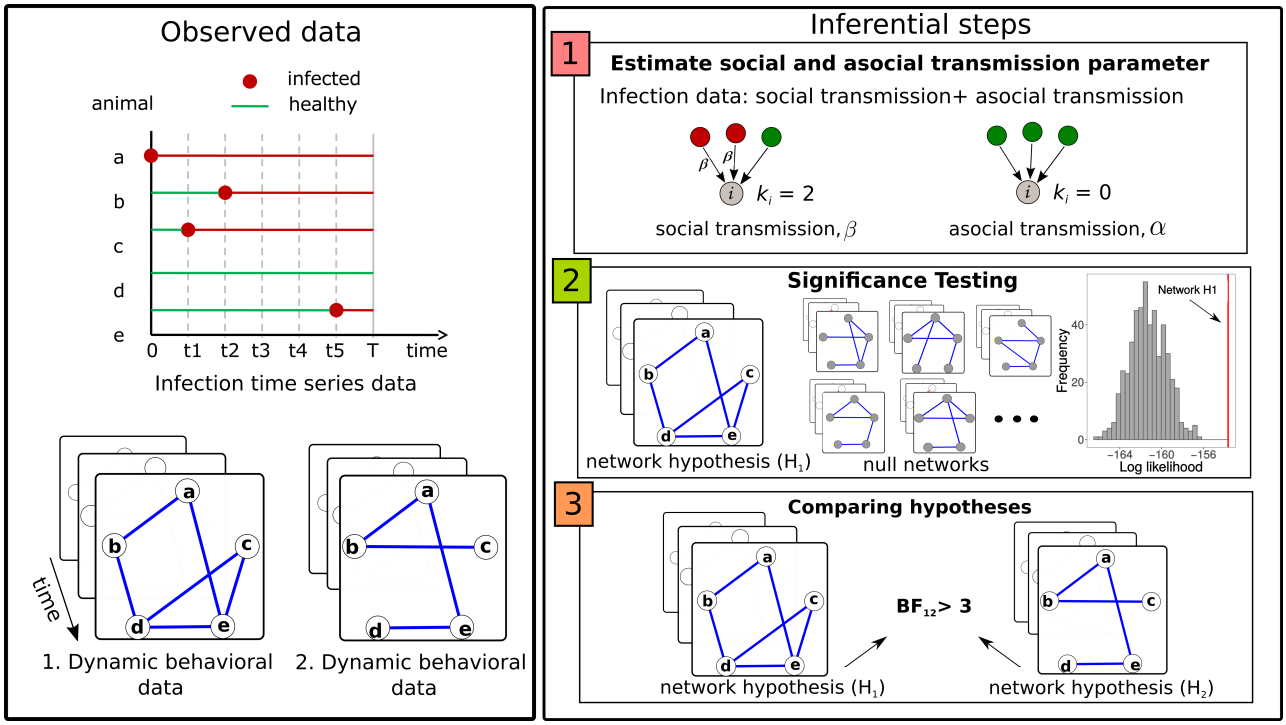


Figure 1. Visualization of the steps of our algorithm. **Observed data**: INoDS utilizes the observed infection time-series and dynamic (or static) interaction data to provide evidence towards contact network hypotheses (or hypothesis) using a three step procedure. **Inferential steps**: In the first step, the tool estimates two parameters for a contact network hypothesis - transmission rate β, which represents the component of infection transmission that is contributed by the network connections, and α that quantifies the component of infection propagation unexplained by the contact network. Significance of β is estimated by comparing the magnitude of social and asocial transmission rate for all observed infection events. Second, 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. Finally, 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.

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 at other geographical locations.

Results

The primary purpose of INoDS is to obtain 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. Evidence is estimated using a three step procedure. First, the tool estimates the unknown social (β) and asocial (α) 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 'missingness' of network hypothesis in explaining certain infection events. In the second step, the predictive power (measured as the likelihood the observed infection time-series data given the network) of network hypothesis is compared against a null expectation. The null expectation is generated by permuting the edge connections of the dynamic network, while controlling for the number of nodes and edges present in the network. A P-value is estimated as the probability of obtaining the observed

predictive power, or something more extreme, if the null expectation were true. 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. 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.

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 (***Molloy and Reed, 1995***)) 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 (***Molloy and Reed, 1995***), for a total of 100 time-steps.

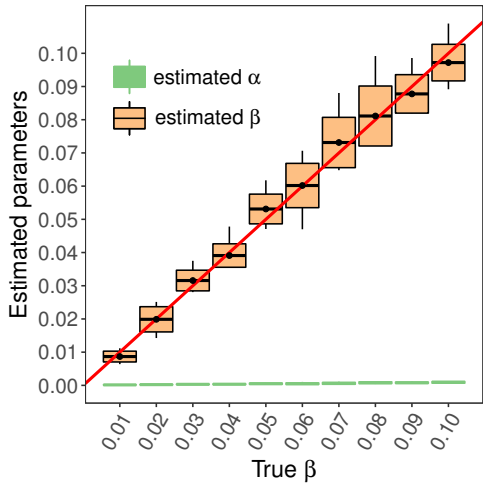


Figure 2. INoDS performance on recovering the social (β) and the asocial (α) transmission parameter of simulated susceptible-infected (SI) model of infectious disease spread. Disease simulations were performed on dynamic networks with 100 nodes, Poisson degree distribution, mean degree of 4, and 100 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 β value estimated by INoDS and the value of β used for disease simulations. Since the simulations were performed on a known synthetic network, the expected value of asocial transmission parameter is zero.

Accuracy of INoDS in estimating the transmission parameters

To test the ability of INoDS to precisely estimate the social transmission parameter β, we performed 10 independent simulations of SI disease spread with transmission rates (β) ranging from 0.01 to 0.1. Model accuracy was determined by comparing the estimated value of transmission parameter β to the value of β 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 α, from the expected value of zero. In Figure 2, we show that INoDS is able to recover the true value of β and α, and that the accuracy of INoDS is independent to the spreading rate of the pathogen.

Robustness of INoDS and previous approaches to missing data

Model robustness was evaluated 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. Model robustness was determined as the proportion of times an approach

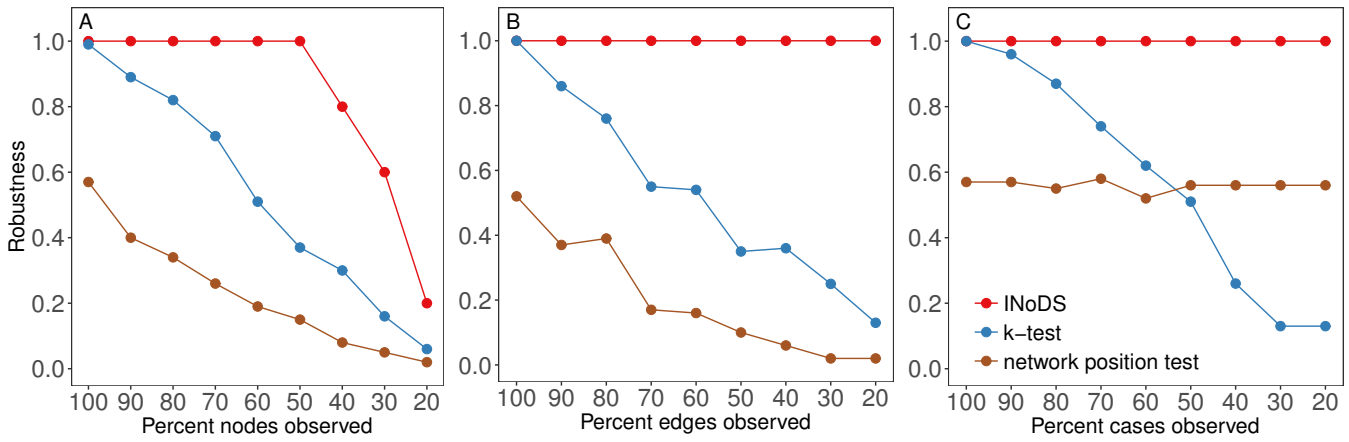


Figure 3. Robustness of INoDS, k-test and network position test to three common forms of missing data - missing nodes, missing edges and missing infection cases. Robustness is calculated as 1 - false negative rate of the test. To calculate the false negative rate, null expectation in INoDS and network position test is generated by permuting the edge connections of the observed networks, creating an ensemble of null networks. On the other hand, in k-test the location of infection cases within the observed network are permuted, creating a permuted distribution of k-statistic.

correctly assigned the observed network with a high statistical power ($P < 0.05$). We compared the robustness of INoDS to two previous approaches – k-test procedure and network position test. 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 (***VanderWaal et al., 2016***). 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 3 summarizes the results. INoDS correctly assigns the observed network with a higher statistical power across a wide range of incomplete sampling scenarios. The tool is least sensitive to missing edges in the observed network and unobserved infection cases. For data with incomplete sampling of nodes, the statistical power of INoDS remains close to one when at least 50% of nodes were observed in the network. We also find that the robustness of this approach is relatively unaffected by the edge density of the underlying contact network (Appendix Fig 1). On the other hand, k-test and network position test were sensitive to all types of missing data. Robustness of k-test procedure rapidly declined with any increase in missing data on the nodes, edges of the observed contact network, and infected cases. Out of the three approaches, network position test proved to be the least reliable as its false negative rate was high even when no data were missing.

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 ***Otterstatter and Thomson (2007)*** 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* (***Bull et al., 2012***).

Determining the role of edge weights in contact network model: 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 foragers. 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. ***Otterstatter and Thomson*** (***2007***) 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 two 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?

The contact networks collected in the previous study were fully connected. As fully connected network rarely describe the dynamics of infection spread, we created sparser contact networks by sequentially removing a certain percentage of weak weighted edges. 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. In addition, four types of contact networks with different edge-weights were constructed - networks where edges were weighted with respect to frequency of contacts, edge weighted by duration of contacts, and edges with no edge weights (i.e., binary edges). Since the technique of creating sparser networks involved removal of weak weighted edges, two types of binary contact networks were constructed - binary networks whose edge connections were the same as frequency weighted contact networks, and binary networks whose edge connections were similar to duration weighted contact networks.

Figure 4 A-D shows the parameter estimates for the four types of contact network hypotheses at different edge weight thresholds. Generally, the rate of infection transfer through social connections was higher than asocial transmission for contact networks where less than 20% of weakest edge weights were removed (dark blue bars in Figure 4). In addition, frequency weighted contact networks (with 5-10% of weakest edge weights removed) were associated with highest Bayesian evidence in all of the analyzed bee colonies (Figure 4E, Appendix figure 3).

Identifying contact network model using 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* (***Bull et al., 2012***). The two sites consisted of 43 and 44 individuals respectively, and these represented close to 100% of all adult lizards 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. Following ***Bull et al.*** (***2012***), we constructed proximity networks 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. Salmonella infection were monitored at the two populations using single cloacal swabs on each animal once every 14 days. Consequently, the disease data in this system represents the timing of *diagnoses* instead of actual infection timings of lizards in the two populations. Since lizards are known to recover from Salmonella infection, and eventually get reinfected, we assumed the infection to follow a SIS (susceiptive-infected-susceptible) disease model.

To test the hypothesis that the proximity interactions of sleepy lizards are predictive of Salmonella spread, we formed two hypothesis of contact network models - i) proximity networks with binary edges, and ii) proximity networks where edges were weighted by frequency (i.e., the total number) of interactions between individuals. Employing Bayesian data augmentation method, we considered the actual infection timings to be unobserved parameters, which was sampled for each contact

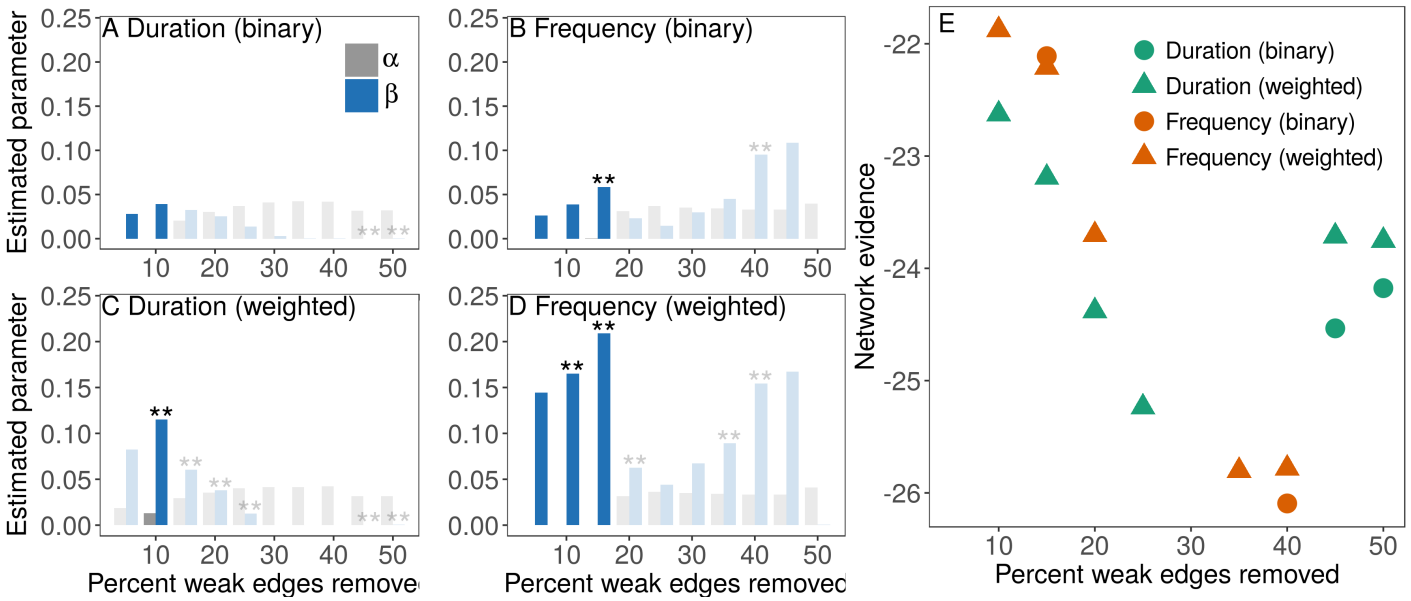


Figure 4. Identifying the contact network of *Crithidia* spread in bumble bee colony (colony UN2). (**A-D**) Edges in the contact network models represent physical interaction between the bees. Since the networks were fully connected, a series of filtered contact networks were constructed by removing 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 - frequency and duration of contacts. In addition, across all ranges of percent weak edges removed, the two types of weighted network were converted to binary networks. The results shown are estimated values of social transmission parameter β, and estimated values of asocial transmission parameter α, for the different contact network hypotheses. Asterisks (**) indicate significant predictive power of the contact network hypothesis as compared to an ensemble of permuted networks. The darker blue color corresponds to the networks, where the social transmission rate was significantly higher than the asocial transmission. (**E**) Log Bayesian (marginal) evidence of the contact networks with high predictive power (i.e., networks marked with asterisks in A-D)

network hypothesis (along with transmission parameters) using data on diagnosis timings. Figure 5 summarizes the results of significance testing (step 2 of INoDS analysis) for site 1. We found that the likelihood of Salmonella infection spreading through the weighted proximity network was significantly greater than the null expectation (Figure 5b. In contrast, the predictive power of binary proximity network was statistically insignificant (P = 0.602). Bayes factor of the weighted network versus the binary network was 817.09 which indicates that the occurrence of repeated contacts between two individuals, rather than just the presence of a contact, is important for Salmonella transmission. Analysis of the two network hypotheses at the second site yielded similar results, although the binary network at the site was also found to be statistically significant (albeit with low predictive power as compared to the weighted network, Appendix 1 Figure 1).

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 (***Jennions and Møller, 2003***; ***Johnson et al., 2015***). 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

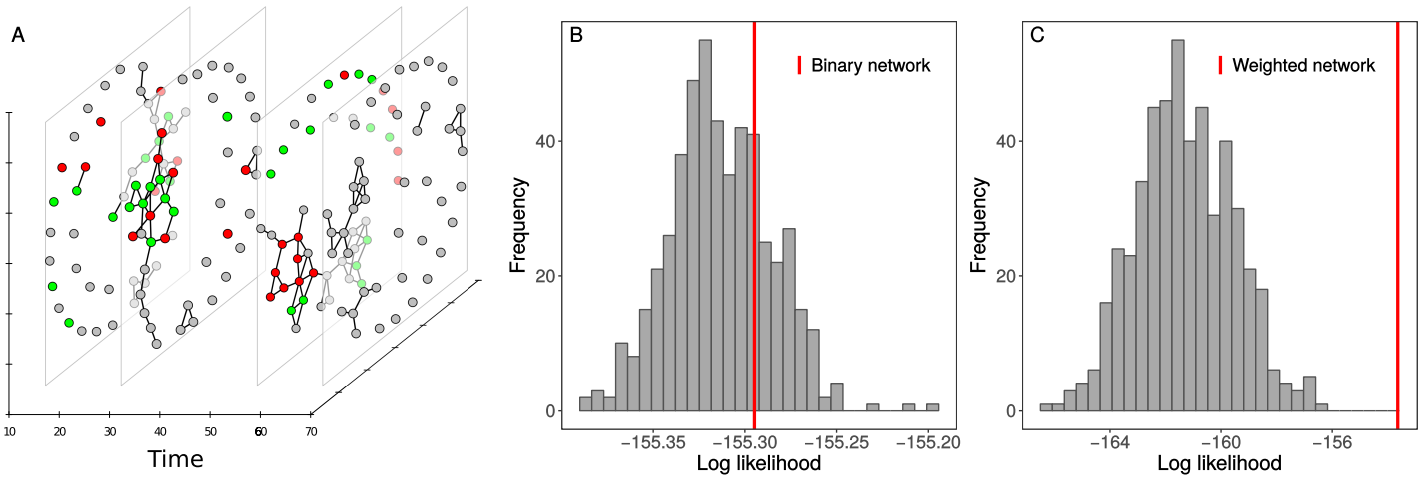


Figure 5. Identifying the contact network of Salmonella spread in Australian sleepy lizards. (A) Dynamic network of proximity interactions between 43 lizards in a population recorded for a total duration of 70 days. Each temporal slice summarizes interaction within a day. Node colors denote the infection status confirmed by laboratory diagnosis. 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 time-point. We hypothesized that proximity network with edges that were either binary or weighted with respect to the frequency of interactions could potentially explain the observed spread of *Salmonella* in the population. We found that (B) the predictive power of binary networks was statistically insignificant (P = 0.602, log-Bayesian evidence = -1039.89), whereas (C) proximity networks with frequency weighted edges demonstrated a significantly greater predictive power that a random expectation (P < 0.001, log-Bayesian evidence = -222.80). The log Bayes factor of the weighted contact network versus binary contact network, is 817.09, which is a decisive evidence towards the frequency weighted proximity network model.

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, β). Unlike previous approaches, INoDS 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 (***Welch et al., 2011***; ***Stack et al., 2013***). The proposed tool in this study can be used to tackle several fundamental challenges in the field of infectious disease modeling (***Eames et al., 2015***). 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 example. For example, spatial proximity, home-range overlap or asynchronous refuge use are commonly used as a proxy of contact in wild animal populations (***Godfrey et al., 2010***; ***Leu et al., 2010***; ***Sah et al., 2016***). 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 (***Pellis et al., 2014***). Through model selection of contact networks with similar edge connects 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 should be weighted with respect to frequency, rather than the duration, of contacts between individual. Our results therefore support the original finding of the study (***Otterstatter and Thomson, 2007***), where individual's risk of infection was found to be correlated with their contact rate with infected nest-mates. However, our results show that edges below a certain edge weight 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 through contaminated faecal scats rather than indirect environmental transmission (***Bull et al., 2012***).

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 underling 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, so 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.

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, with one node pair per line. The software is implemented in Python, is platform independent, and is freely available at (link - to be updated).

INoDS formulation

We assume that at each instance the infection potential for an individual *i* depends on two process - (a) social transmission β, that is captured by the edge connections in the contact network hypothesis, and (b) asocial transmission α, that represents those infection events that are not explained by the edge connections. The infection potential, $\lambda_i(t_i)$, of individual i at the time of acquisition of infection t_i is then calculated as:

$$\lambda_i(t_i) = 1 - \exp\{-\beta k_i(t_i - 1) - \alpha\} \quad (1)$$

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.

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

$$\log(D|H_A, \beta, \alpha) = \sum^{n} \log[\lambda_n(t_n)] + \sum^{t} \left(\sum^{m} \log[1 - \lambda_m(t)] \right) \quad (2)$$

where t_n is the time of infection of individual n. The first part of equation 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.

Data augmentation for unknown infection and recovery time

Calculation of network likelihood using equation 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 popualations 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 ***Tanner and Wong*** (***1987***). 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 $(L_i, d_i]$, 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 $t_i \in \{g_i(t-1) > 0, L_i < t_i \leq d_i\}$, where $g_i(t-1)$ is the degree (number of contacts) of individual i at time $t-1$. 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 (***Robert and Casella, 2004***). 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 from the set of possible recovery time-points using the inverse transform sampling technique.

Estimating the transmission parameters

We adopt a Bayesian approach to estimate the social β and asocial α 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 of for a set Θ of parameters can be written as

$$P(\Theta|D,H) = \frac{\mathcal{L}(D|H,\Theta)\mathcal{P}(\Theta|H)}{\mathcal{E}(D|H)} \propto \mathcal{L}(D|H,\Theta)\mathcal{P}(\Theta|H) \tag{3}$$

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. To obtain the posterior distributions, we use *emcee* package in Python (***Foreman-Mackey et al., 2013***).

Testing the significance of social transmission parameter

To test for the significance of the social transmission parameter (β), the social transmission rate is compared to the asocial transmission rate for each infected individual. The estimated value of β is considered to be statistically significant when the asocial component of infection potential (=α) is greater than the social component (=$\beta k_i(t_i - 1)$) in fewer than 5% of the infection cases.

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.

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:

$$P(D|H,\Theta) = \int \mathcal{P}(\Theta|H)\mathcal{L}(D|H,\Theta)d\Theta \tag{4}$$

Since it is difficult to integrate Eq.4 numerically, we estimate the marginal likelihood of network models using thermodynamic integration, or path sampling (***Lartillot and Philippe, 2006***) 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:

$$log(B_{21}) = log(P(D|H_2)) - log(P(D|H_1)) \tag{5}$$

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) (***Kass and Raftery, 1995***).

Acknowledgments

This work was supported by the National Science Foundation Ecology and Evolution of Infectious Diseases Grant 1216054.

References

Aiello CM, Nussear KE, Esque TC, Emblidge PG, Sah P, Bansal S, et al. Host contact and shedding patterns clarify variation in pathogen exposure and transmission in threatened tortoise *Gopherus agassizii*: implications for disease modeling and management. Journal of Animal Ecology. 2016; p. n/a–n/a. doi: 10.1111/1365-2656.12511.

Bansal S, Grenfell BT, Meyers LA. When individual behaviour matters: homogeneous and network models in epidemiology. Journal of the Royal Society, Interface / the Royal Society. 2007 oct; 4(16):879–91. doi: 10.1098/rsif.2007.1100.

Bansal S, Read J, Pourbohloul B, Meyers LA. The dynamic nature of contact networks in infectious disease epidemiology. Journal of Biological Dynamics. 2010; 4(5):478–489. doi: 10.1080/17513758.2010.503376.

Bull CM, Godfrey SS, Gordon DM. Social networks and the spread of Salmonella in a sleepy lizard population. Molecular Ecology. 2012; 21(17):4386–4392. doi: 10.1111/j.1365-294X.2012.05653.x.

Craft ME, Caillaud D. Network models: An underutilized tool in wildlife epidemiology? Interdisciplinary Perspectives on Infectious Diseases. 2011 jan; 2011:676949. doi: 10.1155/2011/676949.

Danon L, Ford AP, House T, Jewell CP, Keeling MJ, Roberts GO, et al. Networks and the epidemiology of infectious disease. Interdisciplinary perspectives on infectious diseases. 2011 jan; 2011:284909. doi: 10.1155/2011/284909.

Eames K, Bansal S, Frost S, Riley S. Six challenges in measuring contact networks for use in modelling. Epidemics. 2015; 10:72–77. doi: 10.1016/j.epidem.2014.08.006.

Farine D. The dynamics of transmission and the dynamics of networks. Journal of Animal Ecology. 2017; 86(3):415–418. doi: 10.1111/1365-2656.12659.

Fefferman NH, Ng KL. How disease models in static networks can fail to approximate disease in dynamic networks. Physical Review E - Statistical, Nonlinear, and Soft Matter Physics. 2007; 76(3):1–11. doi: 10.1103/PhysRevE.76.031919.

Fenner AL, Godfrey SS, Michael Bull C. Using social networks to deduce whether residents or dispersers spread parasites in a lizard population. Journal of Animal Ecology. 2011; 80(4):835–843. doi: 10.1111/j.1365-2656.2011.01825.x.

Foreman-Mackey D, Hogg DW, Lang D, Goodman J. emcee: The MCMC Hammer. Publications of the Astronomical Society of the Pacific. 2013; 125(925):306–312. doi: 10.1086/670067.

Godfrey SS. Networks and the ecology of parasite transmission: A framework for wildlife parasitology. International journal for parasitology Parasites and wildlife. 2013 dec; 2(1):235–245. doi: 10.1016/j.ijppaw.2013.09.001.

Godfrey SS, Bull CM, James R, Murray K. Network structure and parasite transmission in a group living lizard, the gidgee skink, Egernia stokesii. Behavioral Ecology and Sociobiology. 2009; 63(7):1045–1056. doi: 10.1007/s00265-009-0730-9.

Godfrey SS, Moore JA, Nelson NJ, Bull CM. Social network structure and parasite infection patterns in a territorial reptile, the tuatara (Sphenodon punctatus). International Journal for Parasitology. 2010; 40(13):1575–1585. doi: 10.1016/j.ijpara.2010.06.002.

Hamede RK, Bashford J, McCallum H, Jones M. 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. 2009 nov; 12(11):1147–57. doi: 10.1111/j.1461-0248.2009.01370.x.

Jennions M, Møller A. A survey of the statistical power of research in behavioral ecology and animal behavior. Behavioral Ecology. 2003; 14(3):438–445. doi: 10.1093/beheco/14.3.438.

Johnson PCD, Barry SJE, Ferguson HM, Müller P. Power analysis for generalized linear mixed models in ecology and evolution. Methods in Ecology and Evolution. 2015; 6(2):133–142. doi: 10.1111/2041-210X.12306.

Kass R, Raftery A. Bayes Factors. Journal of the American Statistical Association. 1995; 90(430):773–795. doi: 10.1080/01621459.1995.10476572.

Lartillot N, Philippe H. Computing Bayes Factors Using Thermodynamic Integration. Systematic Biology. 2006; 55(2):195–207. doi: 10.1080/10635150500433722.

Leu ST, Kappeler PM, Bull CM. Refuge sharing network predicts ectoparasite load in a lizard. Behavioral ecology and sociobiology. 2010 sep; 64(9):1495–1503. doi: 10.1007/s00265-010-0964-6.

MacIntosh AJJ, Jacobs A, Garcia C, Shimizu K, Mouri K, Huffman Ma, et al. Monkeys in the middle: parasite transmission through the social network of a wild primate. PloS one. 2012 jan; 7(12):e51144. doi: 10.1371/journal.pone.0051144.

Molloy M, Reed B. A Critical Point for Random Graphs With a Given Degree Sequence. Random Structures and Algorithms. 1995; 6(2-3):161–180.

Otterstatter MC, Thomson JD. Contact networks and transmission of an intestinal pathogen in bumble bee (Bombus impatiens) colonies. Oecologia. 2007 nov; 154(2):411–21. doi: 10.1007/s00442-007-0834-8.

Pellis L, Ball F, Bansal S, Eames K, House T, Isham V, et al. Eight challenges for network epidemic models. Epidemics. 2014; 10:58–62. doi: 10.1016/j.epidem.2014.07.003.

Robert C, Casella G. Monte Carlo Statistical Methods. Springer Texts in Statistics; 2004.

Sah P, Leu ST, Cross PC, Hudson PJ, Bansal S. Unraveling the disease consequences and mechanisms of modular structure in animal social networks. Proceedings of the National Academy of Sciences of the United States of America. 2017 apr; 114(16):4165–4170. doi: 10.1073/pnas.1613616114.

Sah P, Nussear KE, Esque TC, Aiello CM, Hudson PJ, Bansal S. Inferring social structure and its drivers from refuge use in the desert tortoise, a relatively solitary species. Behavioral Ecology and Sociobiology. 2016; p. 1–13. doi: 10.1007/s00265-016-2136-9.

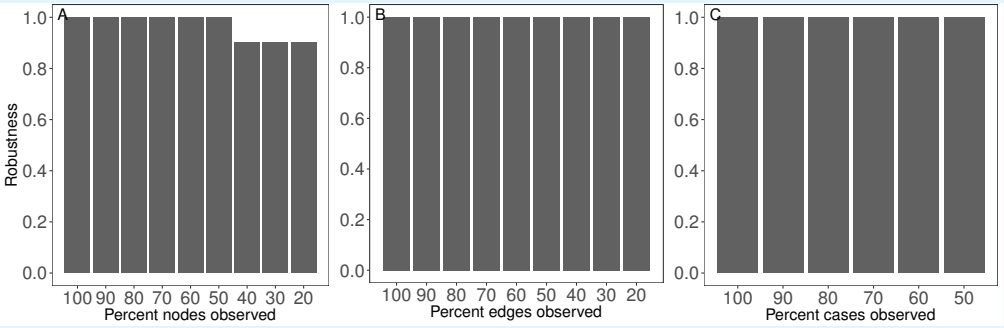
Stack JC, Bansal S, Anil Kumar VS, Grenfell B. Inferring population-level contact heterogeneity from common epidemic data. Journal of the Royal Society, Interface / the Royal Society. 2013; 10(78):20120578. doi: 10.1098/rsif.2012.0578.

Tanner MA, Wong WH. The Calculation of Posterior Distributions by Data Augmentation: Rejoinder. Journal of the American Statistical Association. 1987; 82(398):548–550. doi: 10.2307/2289463.

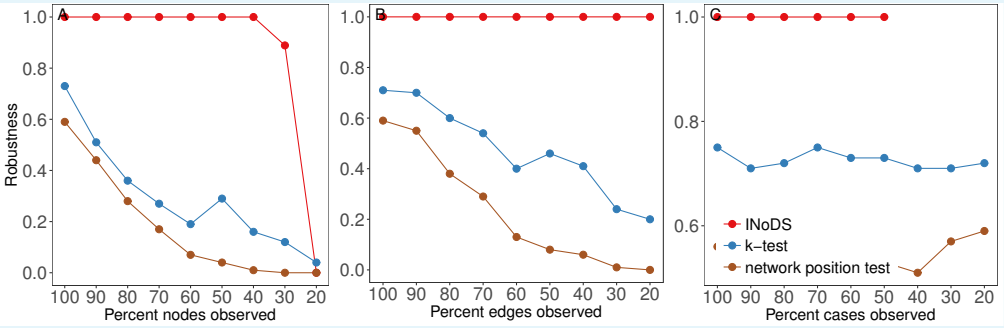
VanderWaal K, Enns EA, Picasso C, Packer C, Craft ME. Evaluating empirical contact networks as potential transmission pathways for infectious diseases. Journal of The Royal Society Interface. 2016; 13(121):20160166. doi: 10.1098/rsif.2016.0166.

Welch D, Bansal S, Hunter DR. Statistical inference to advance network models in epidemiology. Epidemics. 2011; 3(1):38–45. doi: 10.1016/j.epidem.2011.01.002.

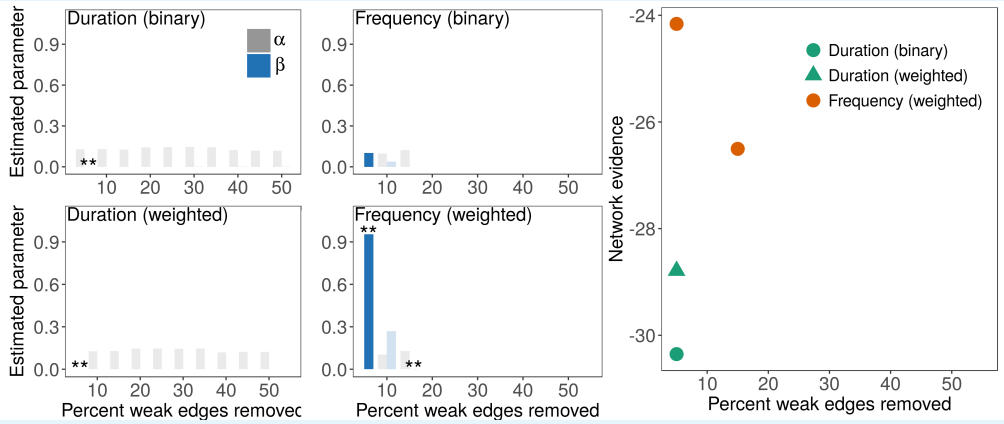
Appendix 1



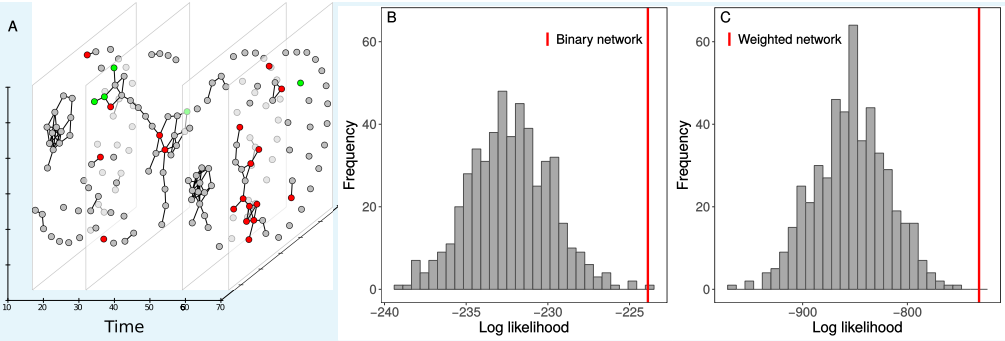
Appendix 1 Figure 1. Robustness analysis of INoDS on sparse dynamic network with average degree of 2.



Appendix 1 Figure 2. Robustness analysis of INoDS, k-test, and network position test performed on simulated spread of pathogen (SI disease model, transmissibility = 0.03) on *static* network (instead of dynamic network in Fig. 3) with 100 nodes and average degree of 3.



Appendix 1 Figure 3. Identifying the contact network of *Crithidia* spread in bumble bee colony (colony UN1). (**A-D**) 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 - frequency and duration of contacts. In addition, across all ranges of percent weak edges removed, the two types of weighted network were converted to binary networks. The results shown are estimated values of social transmission parameter β, and estimated values of asocial transmission parameter α, for the different contact network hypotheses. The darker blue color corresponds to the networks, where the social transmission parameter, β was significantly higher than the asocial transmission parameter, α. Asterisks (**) indicate significant predictive power of the contact network hypothesis as compared to an ensemble of permuted networks. (**E**) Log Bayesian (marginal) evidence of the contact networks with high predictive power (i.e., networks marked with asterisks in A-D)



Appendix 1 Figure 4. Identifying the contact network of Salmonella spread Australian sleepy lizards (second site). (A) Dynamic network of proximity interactions between 43 lizards in a population recorded for a total duration of 70 days. Each temporal slice summarizes interaction within a day. Node colors denote the infection status confirmed by laboratory diagnosis. 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 time-point. We hypothesized that proximity network with edges that were either binary or weighted with respect to the frequency of interactions could potentially explain the observed spread of *Salmonella* in the population. The predictive power of both (B) binary network (P < 0.001, log-Bayesian evidence = -766.50) and (C) frequency weighted network (P < 0.001, log-Bayesian evidence = -255.85) was statistically significant. However, the log Bayes factor of the weighted contact network versus binary contact network, is 510.65, which is a decisive evidence towards the frequency weighted proximity network model.