The application of statistical network models in disease research

Running title: Statistical network models and disease

Matthew J. Silk[1*], Darren P. Croft[2], Richard J. Delahay[3], David J. Hodgson[4], Nicola Weber[4], Mike Boots[4,5] and Robbie A. McDonald[1*]

[1] Environment and Sustainability Institute, University of Exeter, Penryn, Cornwall, UK

[2] Centre for Research in Animal Behaviour, University of Exeter, Exeter, UK

[3] National Wildlife Management Centre, Animal and Plant Health Agency, Woodchester Park, Gloucestershire, GL10 3UJ, UK

[4] Centre for Ecology and Conservation, University of Exeter, Penryn, Cornwall, UK.

[5] Department of Integrative Biology, University of California, Berkeley, California, USA.

*corresponding authors:

MJS: Environment and Sustainability Institute, University of Exeter, Penryn, Cornwall, UK. matthewsilk@outlook.com

RAM: Environment and Sustainability Institute, University of Exeter, Penryn, Cornwall, UK. r.mcdonald@exeter.ac.uk. +44 (0)1326 255720

Main text word count: 6913

Author contributions:

MJS conceived the manuscript, with all authors contributing to developing its full scope. MJS conducted example analyses. NW, RJD and RAM were involved in the original data collection. All authors contributed to drafts and gave final approval for publication

Abstract

1. Host social structure is fundamental to how infections spread and persist and so the statistical modelling of static and dynamic social networks provides an invaluable tool to parameterise realistic epidemiological models.
2. We present a practical guide to the application of network modelling frameworks for hypothesis testing related to social interactions and epidemiology, illustrating some approaches with worked examples using data from a population of wild European badgers *Meles meles* naturally infected with bovine tuberculosis.
3. Different empirical network datasets generate particular statistical issues related to non-independence and sampling constraints. We therefore discuss the strengths and weaknesses of modelling approaches for different types of network data and for answering different questions relating to disease transmission.
4. We argue that statistical modelling frameworks designed specifically for network analysis offer great potential in directly relating network structure to infection. They have the potential to be powerful tools in analysing empirical contact data used in epidemiological studies, but remain untested for use in networks of spatio-temporal associations.
5. As a result, we argue that developments in the statistical analysis of empirical contact data are critical given the ready availability of dynamic network data from bio-logging studies. Further, we encourage improved integration of statistical network approaches into epidemiological research to facilitate the generation of novel modelling frameworks and help extend our understanding of disease transmission in natural populations.

Key words: contact network, epidemiology, temporal network autocorrelation model, exponential random graph model, network-based diffusion analysis, stochastic actor-oriented model, relational event model

Introduction

Direct contact is critical to the transmission of many of the most important infectious diseases and so an understanding of contact networks is integral to the epidemiology of many parasites and pathogens (Keeling & Eames 2005; Read *et al.* 2008; Danon *et al.* 2011; Craft 2015). Populations are not completely mixed and significant population structure arises from spatial (Webb *et al.* 2007a,b) and social interactions. A growing number of empirical studies in humans (Rohani *et al.* 2010; Stehlé *et al.* 2011; Eames *et al.* 2012) and non-human animals (reviewed in Craft 2015; White *et al.* 2015) have found important effects of social network structure on epidemiology, both at an individual- and a population-level. As a result, many epidemiological models now incorporate some concept of non-random social structure that has important consequences for understanding the spread of infections (Keeling & Eames 2005; Lloyd-Smith *et al.* 2005; Craft 2015).

It may also be important to consider networks as dynamic, rather than static, structures, with changes affecting transmission over longer timescales, particularly in endemic diseases (Funk *et al.* 2010, Ezenwa *et al.* 2016, Silk *et al.* 2017). Not only will the temporal structure of interactions have a direct influence on transmission opportunities, but social behaviour may change in response to infection, including both the behaviour of the infected or diseased individual and the response of other individuals towards it (Bansal *et al.* 2010; Croft *et al.* 2011a). Further, these changes in behaviour have been shown to alter contact network structure, with implications for transmission (Tunc & Shaw 2014; Lopes *et al.* 2016). Therefore, accounting for the dynamics of network structure and of infection is key to improving our understanding of disease spread and control in many systems (fig. 1; Bansal *et al.* 2010; Wang *et al.* 2010).

An increasing number of theoretical studies have modelled disease on dynamic networks (e.g. Eames *et al.* 2012; Tunc & Shaw 2014), however there has been relatively little use of empirical data to explore this topic (but see Rohani *et al.* 2010; Reynolds *et al.* 2015; Lopes *et al.* 2016). Using empirical data to test hypotheses about the relationship between sociality and disease (e.g. Drewe

2009; Weber *et al.* 2013) will substantially advance our understanding of the dynamics of infection transmission, and using the outputs of statistical models could help improve the parameterisation of predictive, analytical epidemiological models (Rohani *et al.* 2010; Hamede *et al.* 2012; Reynolds *et al.* 2015). Nevertheless, there are unique problems associated with applying conventional statistical modelling approaches to network datasets (Croft *et al.* 2011b; Farine & Whitehead 2015). First, and perhaps most importantly, social networks recognise the influence of community members on each other, causing non-independence that must be accounted for statistically. Second, social networks are rarely described completely. The impact of sampling process on network parameters should be accounted for in statistical models. This is a particular problem if there is variation among individuals in the completeness of sampling. While this can be an issue for interaction-based networks (here defined as networks constructed from biologically-relevant interactions), it is especially problematic in association-based networks (here defined as networks constructed by connecting individuals that have shared particular groups or spatio-temporal colocations rather than directly to each other), where the extent of sampling is harder to directly assess.

A range of modelling approaches (Table 1), developed within the field of social network analysis, could be applied to study infection in contact networks. These are split broadly into models that continue to use individual traits as a dependent variable while accounting for network structure, and models that use network topology as a dependent variable. The latter could be particularly valuable by directly relating network structure to infection and transmission. Several of these approaches model networks dynamically and offer great potential to improve our understanding of the dynamics of social behaviour and disease. Here we outline these statistical network approaches and provide a guide for how they can best be applied to test a variety of hypotheses related to infection in different types of network. For a selection of modelling frameworks, we use example data from a population of European badgers *Meles meles* naturally infected with bovine tuberculosis (bTB), to illustrate how the approaches can be applied.

Models for static networks

General and generalised linear models, and network autocorrelation models

Traditional statistical modelling frameworks offer an appealing solution to understanding how infection status and social position co-vary with other individual traits. In particular, the use of generalised linear models (GLMs) and generalised linear mixed models (GLMMs) can help study the relationship between social network position and disease state in the context of other predictor traits (e.g. sex, age, physiological condition), either controlling for these traits or considering interactions with them. However, the non-independence of nodes and edges within a network complicates the use of GLMs and GLMMs (Croft *et al.* 2011b), which assume statistical independence of residuals. Also, association-based networks (especially frequent for animal networks) can lead to further biases introduced by the method of network construction (see Farine and Whitehead 2015 for a simulated example of this).

One approach to adapt these modelling techniques appropriately to network data is to use permutation approaches that rely on randomisations of the network (solving the problem of non-independence) or original datastream (see Croft *et al.* 2011b; Farine & Whitehead 2015). A key difference here emerges between interaction networks and association-based networks. The latter requires permutation of the original datastream, due to additional sampling biases (Farine and Whitehead 2015). For these types of networks, other key considerations in implementing data permutations are likely to be the size of social groups, spatio-temporal constraints on interactions, differences in detectability of particular types of individuals, and differences in the probabilities of interactions within, versus outside, social groups (Croft *et al.* 2011b). While biases generated by incomplete sampling can still occur in interaction-based networks, there is greater potential to control this within a modelling framework. For example, if incomplete sampling results from

differences in the length of time each individual is observed then this can be accounted for within any model used.

The R package **asnipe** (Farine 2013) offers a range of algorithms that shuffle association-based data to randomise such networks. However, it may be most appropriate to design system-specific randomisations. One problem worth highlighting is that using a permutation-based approach to test hypotheses creates confidence intervals around the null hypothesis rather than the estimated parameter. The development of approaches that generate uncertainty around observed network data would be highly beneficial in this regard. One example of this idea is provided by Farine and Strandburg-Peshkin (2015), who created probability distributions of edge weights using Bayesian inference. If GLM or GLMM analyses are completed within a Bayesian framework then this sort of uncertainty can be incorporated into the final analysis

An alternative approach that can be used for interaction- or contact-based networks is to incorporate network autocorrelation into the model within a GLM or GLMM framework to address the issue of covariance driven by network structure. This can be achieved using the package **tnam** incorporated within the **xergm suite of** packages (Leifeld *et al.* 2016), or the function `lnam()` in the package **sna** (Butts 2014) in R. The former is discussed here as it has more comprehensive provisions for dependency structures and can incorporate non-Gaussian error distributions. Models constructed using `tnam()` offer a variety of user-defined dependency terms that control for the expectation that individuals may influence other individuals they interact with within a network (see https://cran.r-project.org/web/packages/tnam/tnam.pdf). For example, the `weightlag()` or `netlag()` terms can incorporate autocovariance related to network distance or the `attribsim()` can incorporate autocovariance related to shared attribute values such as group membership. These functions can incorporate additional arguments to make dependency functions more complex. For example, the `netlag()` term can include a number of network steps over which autocovariance may be expected and a mathematical description of the decay. A potential

disadvantage here is that dependency structures are defined by the user, and it is necessary for them to argue that the dependencies incorporated are appropriate and sufficient for the data in question (there is no goodness of fit test that allows this to be tested within the model). As well as incorporating these autocorrelation terms, network autocorrelation models (NAMs) can fit effects of nodal covariates that are either individual-level network metrics (e.g. centrality metrics, clustering coefficient) or exogenous to the network (e.g. sex, age etc.), and the interactions between them (see https://cran.r-project.org/web/packages/tnam/tnam.pdf). There are some potential issues with negatively-biased parameter estimates for `netlag()` terms that should be considered when interpreting autocovariance terms in these models (Mizruchi and Neuman 2008, Neuman and Mizruchi 2010), although these are typically only problematic in high-density networks.

Network autocorrelation model for bTB infection in badgers

We provide an example of a NAM using our badger data in the supplementary material, in which we model bTB infection status as a function of sex, age and flow centrality while accounting for autocovariance among neighbouring individuals in the network. The results are presented in Table S1. This modelling approach finds a positive effect of between-group flow centrality on the probability of bTB infection, as expected from the results of Weber *et al.* (2013). We also found a strong positive correlation between within-group eigenvector centrality and bTB infection, which is of interest as this was not a metric considered by Weber *et al*. (2013). The model also revealed a weak effect of increasing within-group degree on the probability of infection but we would encourage a tentative interpretation of this given the marginal effect and as no attempt has been made to control for the duration that individuals were monitored in our example analysis. These effects of centrality occur independently of differences associated with age class (adults being more likely to be infected than yearlings) and sex (males being more likely to be infected than females). Individuals were also less likely to be infected if their interactions were biased towards infected, not

uninfected, individuals (the `weightlag()` term). Two phenomena are likely to contribute to this seemingly counter-intuitive finding. First, test positive individuals were considered to be infected (test positive by serology or Interferon Gamma Release Assay; see Weber *et al.* 2013) rather than necessarily infectious (test positive by bacterial culture) thus reducing the expectation of positive network covariance in infection. Second, infected individuals were distributed evenly among the badger social groups in the original study, which focussed on a sub-sample of the wider population with high bTB incidence (Fig. 1 in Weber *et al.* 2013).

Partial matrix regressions using Quadratic assignment procedures

Multiple regression quadratic assignment procedures (MRQAP) facilitate multivariate regressions between matrices with complex dependencies by using permutation-based estimates of statistical significance (Cranmer *et al.* 2016, Martin 1999, Dekker *et al.* 2007). Therefore they offer great utility as a tool to explain social network structure using a set of other dyadic relationships. For an ecologist, these are most likely to represent relatedness, some measure of spatial distance, or potentially some measure of difference in individual attributes (e.g. infection status). MRQAP is an accessible method already in use by ecologists. Its direct application to hypotheses related to infection is somewhat limited because it only models dyadic correlations; however, there are some situations where it may be useful. For example, VanderWaal et al. (2013) used MRQAP to compare social networks and transmission networks in giraffes *Giraffa camelopardalis* while controlling for a number of other variables such as spatial overlap. They showed that social network structure better explained transmission network structure than did networks of spatial overlap.

Multiple options are available for calculating MRQAP regressions for network data. Two more familiar options for ecologists are the `netlm()` function in R package **`sna`** (Butts 2014), or

194 the `mrqap.dsp()` and `mrqap.custom.null()` functions in **`asnipe`** (Farine 2013) that enable
195 MRQAP to be used alongside randomisation-based approaches for networks of associations.

196

197 **Exponential random graph models**

198 Exponential random graph models (ERGMs) form a class of statistical models specific to
199 network analysis. They are edge-based models that model the probability (Robins *et al.* 2007; Lusher
200 *et al.* 2013) or weight (Desmarais and Cranmer 2012, Krivitsky 2012, Wilson *et al.* 2017) of each edge
201 as a function of network structure and the characteristics of individuals (nodes) within the network.
202 Local structural configurations can be used alongside nodal or edge covariates to model the pattern
203 of edges observed (see Table 2). ERGMs fit parameters that produce a distribution of networks
204 centered on the observed network (for more details see Lusher *et al.* 2013). Goodness-of-fit of
205 ERGMs can then be assessed by comparing (non-fitted) metrics from the simulated networks with
206 those from the observed network (Lusher *et al.* 2013). The fitting of ERGMs can be complicated by
207 the fact that many parameter combinations can result in model degeneracy (producing model fits
208 that are either very dense or sparse networks), however, this does reduce the likelihood of
209 misspecified models being used. ERGMs are best used with contact or interaction-based data
210 because association- or group-based methods of network construction include uncertainty regarding
211 the true nature of social associations and introduce sampling biases that need to be controlled for
212 (Croft *et al.* 2011b). It may be possible to utilise two-mode exponential random graph models
213 (modelling networks in which edges can only connect between two sets of nodes) for some
214 association-based network data, especially when the links to specific locations are of interest (i.e.
215 modelling what drives any individual's connections to particular locations or groups rather than to
216 each other). In general, however, a restriction to interaction-based networks will not be a major
217 issue in epidemiological research, which typically employs interaction-based networks.

An advantage of ERGMs is the ability to simulate networks based on the parameters for the structural features, and node and edge characteristics included in the observed network with an appropriately fitted model. ERGMs can be a powerful tool for parameterising uncertainty in any epidemiological models constructed (see Welch *et al.* 2011), and this is likely to be especially useful in understanding disease epidemiology, as small differences in network structure have the potential to substantially alter transmission dynamics. This is especially true for studies that use simulation-modelling of the spread of disease across a network (see Reynolds *et al.* 2015). ERGMs also facilitate modelling of social contacts or interactions in response to individual traits, or the properties of dyads (other relationships between individuals such as relatedness). Individual traits (e.g. sex, age, disease state) can be used to explain both the tendency to form connections, and the likelihood of interacting with similar individuals (assortativity). This offers great potential to test hypotheses about the relationship between individual traits, including disease state, and network topology. For example, infected individuals having more interactions than uninfected individuals or tending to interact more frequently with susceptible individuals will increase risk of exposure at a population level. By contrast, assortment among infected individuals would signify that they associate disproportionately and therefore that infection may be socially, and perhaps spatially, restricted in the population. The same argument applies to traits that make individuals more susceptible to infection. Using relatedness as a dyadic variable is a good illustration: related individuals may be more likely to share a genetic susceptibility to some pathogens, so the relationship between the genetic structure and social structure of the population could influence the spatio-temporal distribution of infection.

ERGMs can be constructed using the packages **`ergm`** (Hunter *et al.* 2008; Handcock *et al.* 2015), **`ergm.count`** (Krivitsky 2015) and **`GERGM`** (Denny *et al.* 2016) in R. The package **`ergm.count`** extends ERGMs to Poisson and geometrically distributed edge weights and the package **`GERGM`** generalises ERGMs to all types of weighted network. The latter is a new tool and its use in the type of networks used for epidemiological research is untested. We provide the most

244 relevant terms used in **ergm** and **ergm.count** in Table 2 and a full list of possible terms is included
245 in the help pages for these packages. The range of possible terms is more limited for **GERGM**. The
246 most important terms to include depend on the type of network being used, any structure implicit to
247 it, and the questions being asked (Table 2). R code for an example ERGM is provided in the
248 supplementary material. The *simulate()* function in these packages can then be used to
249 generate new networks based on the modelled parameters to assess goodness of fit or for use in
250 further analysis or network models. We demonstrate its use in the supplementary material.

251

252 ***ERGM to relate bTB infection and network topology in badgers***

253 We provide an example of ERGM in the supplementary information that links bTB infection
254 to increased number of contacts in a badger social network, and to reveal that males tended to
255 interact with more individuals than females (Table S2). By using an ERGM we were able to control
256 for the structure imposed by social groups, and for variation in group size and the number of
257 individuals collared within groups, in the model structure. One might also control for other
258 constraints in the dataset using nodal or dyadic covariates, for example detection biases caused by
259 variation in signal strength in proximity loggers (Drewe *et al.* 2012). We also used our ERGM to
260 simulate badger networks with the same parameters fitted in the model, and show that they are
261 broadly similar to the observed network, albeit not fully capturing the observed network structure
262 (Fig. S1).

263

264 **Latent space network models**

265 Latent space models offer an alternative method to ERGMs for the modelling of relational
266 data, and effectively act as GLMs for edge values that control for network dependence by placing
267 nodes in k-dimensional space according to their social network distance (Cranmer *et al.* 2016).

Covariates can then include relational/dyadic properties (such as relatedness, or differences in a particular attribute) or an attribute of either node represented as a matrix with the same dimensions as the network, meaning the range of nodal and dyadic covariates is very similar to those for ERGMs (Cranmer *et al.* 2016). The potential applications to hypothesis testing in epidemiological studies are therefore broadly similar to ERGMs, but hypotheses about local network dependencies cannot be tested. Further, interpretation of model coefficients can be complicated if the position of nodes in latent space covaries with values of nodal attributes (Cranmer *et al.* 2015).

Latent space models can be fitted in R using the package **`latentnet`** (Krivitsky & Handcock 2008, Krivitsky and Handcock 2015). Latent space models can model weighted edges with a number of pre-defined error distributions. It is possible to use terms from the **`ergm`** package as explanatory variables in latent space models. However, these are limited to the binary variants of model terms, and do not include terms that induce dyadic dependence (such as those incorporating transitivity) as **`latentnet`** only fits models with dyadic independence. The other possible terms that can be included in the model are provided in the **`latentnet`** manual (https://cran.r-project.org/web/packages/ latentnet/latentnet.pdf).

Network-based diffusion analysis

Network-based diffusion analysis (NBDA) compares the likelihood of explaining the spread of a trait through a population for two individual-based models; one assuming purely asocial acquisition of a trait, the other purely social acquisition of a trait (Franz & Nunn 2009). This tests the extent to which social transmission is responsible for explaining the spread of that novel trait through a population. It requires that a single (static) social network and the specific timing of trait acquisition in each individual needs to be known, although this can be order-based or timing-based (Hoppitt *et al.* 2010). Subsequent developments in the models have enabled Bayesian inference

(Nightingale *et al.* 2014). This approach would be particularly valuable in determining the role of contact networks for the transmission of diseases that may have alternative hosts or be spread indirectly via the environment. This is because it tests the hypothesis that a trait spreads through a network, using asocial transmission as the null hypothesis. The use of NBDA in real world populations may be slightly limited, however, by the requirement to know at least the order in which individuals acquired infection.

Lack of data on the order of infection precludes us from providing a badger case study, however R Code to complete NBDA is available in the relevant literature (e.g. Allen *et al.* 2013; Aplin *et al.* 2015) or online (available: http://lalandlab.st-andrews.ac.uk/freeware/).

Models for dynamic networks

Incorporating a dynamic view of population social structure will greatly enhance applications of social networks to epidemiology. Both social structure and infection are dynamic traits that interact at population and individual levels (Fig. 1; White *et al.* 2015). Two categories of approaches have been suggested: a) modelling the changes in a series of aggregated static networks using GLMMs, stochastic actor-oriented models (Snijders *et al.* 2010) and temporal ERGMs (Hanneke *et al.* 2010), or b) using relational event models (Butts 2008) to model temporally-explicit contact data. Both of these approaches, especially the latter, require high resolution temporal data on social interactions (and to capture co-dynamics similar resolution data on infection), and so their use may be limited to exceptionally detailed datasets.

Generalised linear mixed models and temporal network autocorrelation models

Both randomisation-based GLMM and NAM approaches can be used to study a set of aggregated networks or network snapshots with, in the latter case, the models becoming temporal network autocorrelation models (TNAMs). Randomisation-based GLMM approaches can be extended to network snapshots by including individual as a random effect in a model that relates social network position and disease state (alongside other variables of interest). It is also possible to incorporate change in values of network metrics over time as an additional variable to improve the extent to which these models capture the importance of social dynamics. When GLMMs are used to model a temporal series of networks, the simplest way to design appropriate randomisations would be to permute or randomise the network or association data within the sampling period used to construct each network snapshot (Farine & Whitehead 2015).

TNAMs can incorporate temporal autocorrelation by using the `lag` argument for each model term. This is equally applicable to the response variable re-fitted as a time-lagged covariate, e.g. an individual's disease state being dependent on its disease state in preceding time-steps; other covariates, e.g. an individual's disease state depending on body condition at a previous time-step as well as the current one; and network features, e.g. disease state could depend on the disease state of neighbouring individuals in the network at the current and preceding time-steps. For cases in which changes in disease state are regularly observed, this approach offers great potential to better appreciate the temporal scale over which social relationships influence acquisition of infection. The rate of change in observed bTB infection in badgers is too low relative to our one year sample of contact network data for it to be possible to provide a badger example, but the implementation of TNAMs in R (also using **`tnam/xergm`**) is very similar to that of NAMs.

Stochastic actor-oriented models

Stochastic actor-oriented models (SAOMs) use an individual-based approach to model how network structure changes through time, and can link these changes to structural features of the network, individual traits or dyadic covariates (Snijders et al 2010, Fisher *et al.* 2017). Model terms (structural terms, and individual or dyadic covariates) can be used to explain both the rate that an individual has an opportunity to change to its network position (the "rate" function) and the probability that it does so when the opportunity arises (the "objective" function) (Snijders *et al.* 2010; Ripley *et al.* 2011). Both individual and dyadic covariates can remain fixed (e.g. sex in our example) or change over time, but act only as explanatory variables (e.g. bTB infection in our example). Individual traits can also coevolve with network structure and form part of the response.

SAOMs are most appropriate for use with interaction- or contact-based networks, due to the similar constraints described for ERGMs (i.e. the uncertainty over the true nature of interactions and data structure in association-based networks). However, similarly to ERGMs, it is possible to control for structural features in interaction- or contact-based data using covariates e.g. distance effects or shared group effects (Fisher *et al.* 2017). SAOMs can currently model only binary or ordered networks, so are best used in cases where the presence/absence of an edge is more informative than its weight, or when network snapshots are constructed over relatively short time windows (Fisher *et al.* 2017). However, being able to incorporate ordered networks does at least enable relationships of different strengths to be modelled separately (see http://www.stats.ox.ac.uk/~snijders/siena/RscriptSiena_Ordered.R), which may be important for particular diseases or social systems.

A major advantage of using SAOMs is the ability to model the "co-dynamics" of social strategy and infection status. This would enable better understanding of what drives the correlation between network position and infection status, especially important for research on endemic infections. For example, individuals with more contacts may be more at risk of infection, but it is

equally possible that increases in social contacts are caused directly by infection or disease. Additionally, SAOMs enable the modelling of the influence of disease state and other variables (e.g. sex) on both the probability of individuals forming particular interactions and the rate at which they change these interactions. This helps disentangle how different social strategies influence susceptibility to disease. Finally, an extension of the SAOM framework enables a response variable, for example immunity, to be fixed once it is acquired i.e. no return is possible to the original state (Ripley *et al.* 2011; Greenan 2015), and this may facilitate the addition of immunity into hypothesis testing in real world contact networks.

SAOMs are implemented in R using the package **`RSiena`** (Ripley *et al.* 2013). Models are best constructed in a stepwise manner (see supplementary information), starting with basic structural terms and adding in more complex structural terms, and then behavioural terms, once the current model converges and fits the data at each step (Ilany *et al.* 2015; Fisher *et al.* 2017). The data requirements, as well as details on tests for model convergence, goodness of fit and significance, are provided elsewhere (Ripley *et al.* 2011; Ilany *et al.* 2015; Fisher *et al.* 2017). However, we highlight two important considerations of direct relevance to disease research. First, it is possible to include individuals that were not present at all time points by incorporating structural zeroes into the association matrices (Ripley *et al.* 2011), meaning that individuals that enter or leave a population during the study period can be included. Second, if a trait is intended to coevolve with network structure in the model, it must be a binary or ordinal variable. In disease modelling this is likely to be equivalent to classifying individuals as uninfected or infected, or to using numbers that reflect progressive disease states. For example, multiple classes used to describe bTB infection states in European badgers (e.g. Graham *et al.* 2013), could be coded ordinally.

Using a SAOM to examine seasonal changes in badger interactions

We use an SAOM to explore badger social network dynamics from summer through winter, showing that there is no evidence for bTB increasing either the probability of interactions or the rate at which interactions change for a binary network of all interactions (potentially as a result of using a binary contact network, and the reduced subset of individuals included; n=36, c.f. n=51 for the ERGM). However, there are interesting differences in the rate of network change between the sexes, with males changing their interactions faster than females between summer and winter. Differences such as this may provide a behavioural explanation for males being more likely to acquire infection than females in this system (Graham *et al.* 2013). Furthermore, the significant effects of distance between setts and shared group membership reveal the importance of spatial behaviour in structuring the badger social system, and highlight the importance of accounting for data structure when using statistical models in these ways.

Temporal Exponential Random Graph Models

Temporal ERGMs (TERGMs) represent a generalisation of the ERGM framework to a temporal series of static networks (Hanneke *et al*. 2010, Leifeld *et al*. 2015). TERGMs assume that a network in one time-step is dependent on network structure in the preceding time-steps, with the number of previous time-steps used determined by a parameter within the model.

The ability to simulate networks in longitudinal datasets is a particular advantage of using TERGMs. Studies that use network models of disease in animals often encompass change in network structure over time, for example in response to seasonal changes (Reynolds *et al.* 2015). Therefore TERGMs offer an ideal framework to simulate networks into the future, based on a set of network snapshots. In terms of hypothesis testing, the incorporation of temporal dependencies can enable i) the role of disease in network topology to be estimated while accounting for variation in interaction

stability over time or ii) the role of disease state in influencing temporal changes in interactions to be estimated (if disease state of two individuals is included as a dyadic covariate).

TERGMs can be fitted using the package **`btergm`**, part of the **`xergm`** package suite (Leifeld et al. 2016) in R. The TERGM framework can handle changes in network size between time-steps if row or column labels are provided in the matrix. This can be achieved by removing these nodes or by incorporating them as structural zeroes. However, within a time-step, individuals must possess a full set of network information and covariate values. If this is problematic, it is possible to impute values either for covariates or network data (e.g. Koskinen *et al.* 2013). Basic imputation can be done within the **`xergm`** package.

The `btergm()` function enables models containing time dependent covariates (`timecov()` argument) and effects of tie stability (`memory()` argument) and delayed reciprocity (`delrecip()` argument for directed networks) to be fitted alongside conventional ERGM terms (Table 2; Leifeld *et al.* 2015). The parameter *k* defines the number of preceding time-steps which affect the current time-step. It is possible for *k* to take values greater than 1 but as *k* increases the number of time-steps remaining to model reduces, placing a constraint upon the user. The `timecov()` argument enables interactions between dyadic covariates and temporal trends in edge formation (with the exact nature of the temporal trend provided as a function by the researcher) so is likely to be especially useful in understanding differences in interactions linked to infection status. The provision of a user defined temporal pattern of interactions requires some careful thought from the researcher when implementing the model, but provides a more flexible tool for defining temporal change in network structure than available in SAOMs. Further, other dyadic covariates can vary through time if they are provided as a list of matrices. This is likely to be particularly relevant to individual-level variables, such as disease and state, which also vary temporally.

Example TERGMs for badger-TB epidemiology

We provide some basic examples of the fitting of TERGMs to our dataset in the supplementary material using the same subset of data used for the SAOM example. While only using a temporal series of three networks restricted us to simplified model constructs, we show how the different terms can be used to test hypotheses about changes in network structure over time alongside using individual-level covariates. The first example model shows that there is greater stability in badger contact networks than expected by chance (Table S4), while the second shows that there is a decline in the probability of contacts between summer and winter (Table S5). There is no consistent pattern between models for the effects of bTB infection and sex, suggesting the use of binary network data might be limiting the power of detecting these effects. These example models are also used to show how to use goodness-of-fit tests for TERGMs (Fig. S3). For further information we refer readers to Leifeld & Cranmer (2015) and Leifeld *et al.* (2015).

Relational Event Models

Relational event models (REMs) provide a modelling framework capable of analysing data on contacts, interactions or associations that haven't been aggregated, remain temporally-explicit and are instantaneous events without measurable duration (Butts 2008; Tranmer *et al.* 2015). The concept is similar to event models used in survival analysis, and estimates a hazard function for the rate of interaction events conditional on covariates measured on either individuals or events, and also on patterns of these interactions in the past (Tranmer *et al.* 2015). Within a 'relational' framework it possible to additionally estimate coefficients for the influence of network effects on these events such as transitivity – a tendency to interact with '*friends of friends*' (Butts 2008). It is now possible to incorporate a decay function so that events that have happened more recently have a greater effect (Lerner *et al.* 2013). In addition, another recent extension of the REM framework can

be used to make them applicable to two-mode networks (Brandenberger 2016), in which edges can only connect between two independent sets of nodes. This could extend their use to association-based networks in which individuals are connected to particular groups or locations rather than directly to each other.

The potential applications of REMs to wildlife disease research are manifold, especially given the growing number of studies in this field that use temporally explicit data from proximity loggers (e.g. Hamede *et al.* 2009; Cross et al. 2012; Weber *et al.* 2013). This framework could be highly informative in understanding how the acquisition or progression of an infection influences the likelihood of repeat social contacts with uninfected individuals, or the persistence of an individual's social associations (Fig. 1). Additionally, for populations in which social structure represents an important barrier to the spread of infection, REMs would facilitate the modelling of differences between the dynamics of intra-group and inter-group interactions. The temporal structure of inter-group interactions would be expected to have a substantial effect on disease spread and previous interactions within a dyad, especially those in the recent past, could increase the likelihood of further interactions occurring. Finally, differences in these parameters between the sexes or for individuals of different ages might explain patterns of age- or sex-biased infection.

REMs can be fitted in R using the package **`rem`** (Brandenberger 2016) or using the package **`relevent`** (Butts 2008), with prior data manipulation requiring the package **`informR`** (Marcum and Butts 2015). This includes the addition of support constraints (additional binary indicators within the model that restrict which actions or events are possible) that can help account for elements of the study design, and therefore are likely to be particularly beneficial in studies of animals (Tranmer *et al.* 2015). For example, support constraints could inform a model when individuals are collared in a contact network study, or to indicate whether two individuals are on different sides of a geographical barrier (e.g. a river) and therefore unable to interact. Extensions to incorporate weightings on temporal dependencies among events are incorporated in the **`rem`** package.

Choosing a model

With such a wealth of approaches, it may not be immediately clear which offers the most appropriate tool to test a particular hypothesis. In Table 1 we outline the advantages and disadvantages of using all of the modelling frameworks outlined here. In Figure 2 we provide a data- and question-driven approach to selecting the most suitable statistical tool. For further comparisons between statistical models of networks, and guidance to their usage, we refer readers to recent reviews in other subject areas (Hunter *et al.* 2012, Leifeld and Cranmer 2015, Cranmer *et al.* 2016). In addition to using statistical network models, it may also be possible to use statistical models of contact rates to test hypotheses relating disease and social behaviour, especially within social groups (Cross et al. 2012).

There are a few important general rules to consider when selecting a modelling framework. The first of these is how the network data are obtained. Networks constructed using group-based (or association-based) approaches contain data structure and biases that on current knowledge require randomisation-based approaches that employ GLMs or GLMMs. For networks constructed from defined social contacts or interactions, then any approach could be useful depending on the question of interest. If data are temporally explicit (time-ordered) then the use of REMs offers the most powerful analytical approach by facilitating the use of temporal patterns of contacts in addition to their structure. However, these models are complex to construct and so for answering simpler questions it might be appropriate to aggregate data into a temporal series of networks and use simpler approaches. It may even be that for some questions aggregating all network data into a single static network still enables the relevant hypotheses to be tested.

When selecting between network-focussed statistical models - (T)ERGMs, (T)NAMs and SAOMs - a fundamental first consideration is whether the hypotheses being tested are related to properties of

relational data or the properties of nodes. For hypotheses related to network topology, (T)ERGMs and SAOMs are most appropriate, while for nodes (T)NAMs are best (or alternatively GLMMS with randomisations). Many hypotheses revolving around the topic of social behaviour and disease are in fact most suitable for testing using models of network topology . For example, any question asking whether diseased individuals show different patterns of social behaviour to non-diseased individuals, or asking how social behaviour changes as infection state changes are "network topology" questions. (T)NAMs are especially useful in testing hypotheses linking change in infection status to the network position of an individual and the infection status of individuals surrounding it in the network (alongside any other individual-level fixed effects). Thus modelling how network structure influences the probability of acquiring infection should be considered a "node-based" question.

Missing information and hypothesis testing in networks

Many network studies of disease transmission are likely to contain missing information, either because they are based on a sub-sample of the total population or record only a subset of the interactions that occur amongst individuals. Few studies have investigated the impact of missing information on network analysis (but see e.g. Lee *et al.* 2006, Smith & Moody 2013, Silk *et al.* 2015, Smith *et al.* 2017), and none has gone on to test how different types and levels of missing information affect hypothesis testing approaches. As a result, we would currently urge caution in applying these methods where networks are constructed using only a small proportion of individuals within a study population. An alternative option when there are high levels of missing information is to model contact rates independently of network structure, for example the methods outlined in Cross et al. (2012). If statistical network methods are influenced in different ways by the sub-sampling of network data then the choice of model might also depend on the level of sampling in the network of interest. For example, Shalizi and Rinaldo (2013) suggested that an ERGM based on a sampled network is unlikely to reflect population-level parameters, although how this might affect

the testing of hypotheses is unclear. Conversely, Páez *et al.* (2008) found that the power of NAMs to detect network effects remained high until a majority of edge information was missing. Developing an improved understanding of how different modelling approaches are affected by sampling of a network will be a valuable area of future methodological research.

Network approaches and epidemiological modelling

A natural end point of applying social network analytical methods to the study of disease is in helping to construct and parameterise epidemiological models and there are numerous advantages of this approach. First, uncertainty can be incorporated more easily – any estimates for structural effects or individual differences from ERGMs, SAOMs or REMs will include standard errors, which can be included to test the robustness of the conclusions drawn from the model. Second, statistical models (especially ERGMs) facilitate the easy simulation of large number of networks with equivalent expected properties to the observed network, useful for simulation-modelling of disease. Third, the use of dynamic statistical models (SAOMs, temporal ERGMs) makes it easier to incorporate information on network dynamics into any constructed models. For SAOMs in particular, the ability to estimate the co-dynamics of social strategy and disease could have major implications (e.g. the inclusion of avoidance behaviour in epidemiological models: Shaw & Schwartz 2008; Tunc & Shaw 2014). As a result, the incorporation of these statistical network models alongside epidemiological models offers great potential to develop stronger links between empirical data and disease modelling, especially in models of endemic diseases, for which the co-dynamics of social systems and infection are likely to be more important.

Conclusions and future directions

There is considerable scope to extend current modelling frameworks and it would be highly beneficial for epidemiological researchers to become more involved in their continued development. For example, many of these methods are rather poor at dealing with missing data, and integrating elements from Bayesian population models (using state-space/multi-state models to address the issue of missing data and hidden states: Kéry & Schaub 2012) and models of network topology could make substantial advances in dealing with this issue.

Developments in hypothesis testing in networks will enable important progress in understanding the links between individuals, social structure and infection. This is especially true for endemic infections, such as with our worked examples of bTB in badgers, where the longer timescales involved will mean that understanding the dynamic interaction between social behaviour and disease is that much more important. Furthermore, implementing statistical approaches specifically designed to model networks can facilitate more detailed parameterisation of epidemiological models and provide an idea of uncertainty around key parameters. Together this means that statistical models of networks can offer a powerful tool in linking empirical data on population social structures with theoretical models of disease.

Acknowledgements

MS is funded by a NERC standard grant (NE/M004546/1) awarded to RM, DC, DH and MB, with the APHA team at Woodchester Park, UK (lead scientist is RD) as project partners. Data used in example analyses were collected for NW's PhD funded by Defra.

Data accessibility

Full R code for example models are provided in the supplementary information. The data analysed are also provided as supplementary files.

References

Allen, J., Weinrich, M., Hoppitt, W. & Rendell, L. (2013). Network-based diffusion analysis reveals cultural transmission of lobtail feeding in humpback whales. *Science*, **340**, 485–488.

Aplin, L.M., Farine, D.R., Morand-Ferron, J., Cockburn, A., Thornton, A. & Sheldon, B.C. (2015). Experimentally induced innovations lead to persistent culture via conformity in wild birds. *Nature*, **518**, 538–541.

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

Brandenberger L. (2016). Rem: Relational Event Models. R package version 1.1.2

Butts, C.T. (2008). A relational event framework for social action. *Sociological Methodology*, **38**, 155–200.

Butts, C.T. (2014). sna: Tools for Social Network Analysis. R package version 2.3-2. https://CRAN.R-project.org/package=sna

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

Cranmer, S.J., Leifeld P., McClurg, S.M. & Rolfe M. (2016). Navigating the Range of Statistical Tools for Inferential Network Analysis. American Journal of Political Science. DOI: 10.1111/ajps.12263

Croft, D.P., Edenbrow, M., Darden, S.K., Ramnarine, I.W., van Oosterhout, C. & Cable, J. (2011a). Effect of gyrodactylid ectoparasites on host behaviour and social network structure in guppies Poecilia reticulata. *Behavioral Ecology and Sociobiology*, **65**, 2219–2227.

Croft, D.P., Madden, J.R., Franks, D.W. & James, R. (2011b). Hypothesis testing in animal social networks. *Trends in Ecology & Evolution*, **26**, 502–507.

Cross, P.C., Creech, T.G., Ebinger, M.R., Heisey, D.M., Irvine, K.M. & Creel, S. (2012) Wildlife contact analysis: emerging methods, questions, and challenges. Behavioral Ecology and Sociobiology, 66, 1437–1447.

Danon, L., Ford, A.P., House, T., Jewell, C.P., Keeling, M.J., Roberts, G.O., Ross, J. V & Vernon, M.C. (2011). Networks and the epidemiology of infectious disease. *Interdisciplinary perspectives on infectious diseases*, **2011**.

Dekker, D., Krackhardt D., and Snijders, T.A.B.(2007). Sensitivity of MRQAP Tests to Collinearity and Autocorrelation Conditions. *Psychometrika*, **72**, 563–81.

Denny, M.J., Wilson J.D., Cranmer S., Desmarais, B.A. and Bhamidi S. (2016). GERGM: Estimation and Fit Diagnostics for Generalized Exponential Random Graph Models. R package version 0.10.0.

https://CRAN.R-project.org/package=GERGM

Desmarais, B.A. & Cranmer, S.J. (2012) Statistical inference for valued-edge networks: the generalized exponential random graph model. *PloS one*, **7**, e30136.

Doreian, P., Freeman, L.C., White, D.R. & Romney, A.K. (1989) Models of network effects on social actors. Research methods in social network analysis, 295–317.

Drewe, J.A. (2009). Who infects whom? Social networks and tuberculosis transmission in wild meerkats. *Proceedings of the Royal Society of London B: Biological Sciences*, rspb20091775.

Drewe, J.A., Weber, N., Carter, S.P., Bearhop, S., Harrison, X.A., Dall, S.R., McDonald, R.A. & Delahay, R.J. (2012). Performance of proximity loggers in recording intra-and inter-species interactions: a laboratory and field-based validation study. *PloS one*, **7**, e39068–e39068.

Eames, K.T.D., Tilston, N.L., Brooks-Pollock, E. & Edmunds, W.J. (2012). Measured dynamic social contact patterns explain the spread of H1N1v influenza. *PLoS Comput Biol*, **8**, e1002425.

Ezenwa, V.O., Archie, E.A., Craft, M.E., Hawley, D.M., Martin, L.B., Moore, J. & White, L. (2016) Host behaviour–parasite feedback: an essential link between animal behaviour and disease ecology. Proc. R. Soc. B, p. 20153078. The Royal Society.Farine, D.R. (2013). Animal social network inference and permutations for ecologists in R using asnipe. *Methods in Ecology and Evolution*, **4**, 1187–1194.

Farine, D.R. & Strandburg-Peshkin, A. (2015). Estimating uncertainty and reliability of social network data using Bayesian inference. *Open Science*, **2**, 150367.

Farine, D.R. & Whitehead, H. (2015). Constructing, conducting and interpreting animal social network analysis. *Journal of Animal Ecology*, **84**, 1144–1163.

Fisher, D., Ilany, A., Silk, M. and Tregenza T. 2017. Analysing animal social network dynamics: the potential of stochastic actor-oriented models. *Journal of Animal Ecology*, **86**, 202-212.

Franz, M. & Nunn, C.L. (2009). Network-based diffusion analysis: a new method for detecting social learning. *Proceedings of the Royal Society of London B: Biological Sciences*, **276**, 1829–1836.

Funk, S., Salathé, M. & Jansen, V.A.A. (2010) Modelling the influence of human behaviour on the spread of infectious diseases: a review. Journal of the Royal Society Interface, 7, 1247–1256.

Graham, J., Smith, G.C., Delahay, R.J., Bailey, T., McDonald, R.A. & Hodgson, D. (2013). Multi-state modelling reveals sex-dependent transmission, progression and severity of tuberculosis in wild badgers. *Epidemiology and infection*, **141**, 1429–1436.

Greenan, C.C. (2015). Diffusion of innovations in dynamic networks. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, **178**, 147–166.

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.

Handcock, M., Hunter, D., Butts, C., Goodreau, S., Krivitsky, P. and Morris, M. (2015). _ergm: Fit,

650 Simulate and Diagnose Exponential-Family Models for Networks_. The Statnet Project (<URL:
651 http://www.statnet.org>). R package version 3.5.1, <URL: http://CRAN.R-
652 project.org/package=ergm>.

653 Hanneke, S., Fu, W. & Xing, E.P. (2010). Discrete temporal models of social networks. *Electronic*
654 *Journal of Statistics*, **4**, 585–605.

655 Hays, J.C., Kachi, A. and Franzes Jr., R.J. (2010). A spatial model incorporating dynamic, endogenous
656 network interdependence: A political science application. *Statistical Methodlogy,* **7**, 406-428.

657 Hoff, P.D., Raftery A.E., and Handcock, M.S. (2002). Latent Space Approaches to Social Network
658 Analysis. *Journal of the American Statistical Association*, **97**, 1090–98.

659 Hoppitt, W., Kandler, A., Kendal, J.R. & Laland, K.N. (2010). The effect of task structure on diffusion
660 dynamics: Implications for diffusion curve and network-based analyses. *Learning & Behavior*,
661 **38**, 243–251.

662 Hunter, D.R., Handcock, M.S., Butts, C.T., Goodreau, S.M. & Morris, M. (2008). ergm: A package to
663 fit, simulate and diagnose exponential-family models for networks. *Journal of statistical*
664 *software*, **24**, 1-29.

665 Hunter, D.R., Krivitsky, P.N. & Schweinberger, M. (2012) Computational statistical methods for social
666 network models. Journal of Computational and Graphical Statistics, 21, 856–882.Ilany, A.,
667 Booms, A.S. & Holekamp, K.E. (2015). Topological effects of network structure on long-term
668 social network dynamics in a wild mammal. *Ecology letters*, **18**, 687–695.

669 Keeling, M.J. & Eames, K.T.D. (2005). Networks and epidemic models. *Journal of the Royal Society*
670 *Interface*, **2**, 295–307.

671 Kéry, M. & Schaub, M. (2012). *Bayesian population analysis using WinBUGS: a hierarchical*
672 *perspective*. Academic Press.

673 Koskinen, J.H., Robins, G.L., Wang, P. & Pattison, P.E. (2013). Bayesian analysis for partially observed
674 network data, missing ties, attributes and actors. *Social Networks*, **35**, 514–527.

675 Krivitsky, P. (2015). _ergm.count: Fit, Simulate and Diagnose Exponential-Family Models for
676 Networks with Count Edges_. The Statnet Project (<URL: http://www.statnet.org>). R package
677 version 3.2.0, <URL: http://CRAN.R-project.org/package=ergm.count>.

678 Krivitsky, P. & Handcock, M. (2015). latentnet: Latent Position and Cluster Models for Statistical
679 Networks_. The Statnet Project (<URL: http://www.statnet.org>). R package version 2.7.1,
680 <URL: http://CRAN.R-project.org/package=latentnet>.

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

683 Krivitsky, P.N., Mark S. Handcock, M.S., Raftery, A.E. & Hoff, P.D. (2009). Representing degree
684 distributions, clustering, and homophily in social networks with latent cluster random effects
685 models. *Social Networks*, **31**, 204-213.

686 Krivitsky, P.N., & Handcock, M.S. (2008). Fitting position latent cluster models for social networks
687 with latentnet. *Journal of Statistical Software*, **24**.

688 Lee, S.H., Kim, P.-J. & Jeong H. (2006). Statistical properties of sampled networks. *Physical Review E*,
689 **73**, p. 016102.

690 Leenders, R.T.A.J. (2002). Modeling social influence through network autocorrelation: constructing
691 the weight matrix. *Social Networks*, **24**, 21-47.

692 Leifeld, P. & Cranmer, S.J. (2015). A theoretical and empirical comparison of the temporal
693 exponential random graph model and the stochastic actor-oriented model. *arXiv preprint*
694 *arXiv:1506.06696*.

695 Leifeld, P., Cranmer, S.J. & Desmarais, B.A. (2015). Temporal Exponential Random Graph Models
696 with xergm: Estimation and Bootstrap Confidence Intervals. *Journal of Statistical Software*.

697 Leifeld, P., Cranmer, S.J., and Desmarais, B.A. (2016). xergm. Extensions for Exponential Random
698 Graph Models. R package version 1.7.0.Lerner, J., Bussmann, M., Snijders, T.A. & Brandes U.
699 (2013). Modeling frequency and type of interactions in event networks. *Corvinus journal of*
700 *sociology and social policy*, **4**, 3-32.

701 Lloyd-Smith, J.O., Schreiber, S.J., Kopp, P.E. & Getz, W.M. (2005). Superspreading and the effect of
702 individual variation on disease emergence. *Nature*, **438**, 355–359.

703 Lopes, P.C., Block, P. & König, B. (2016). Infection-induced behavioural changes reduce connectivity
704 and the potential for disease spread in wild mice contact networks. *Scientific Reports*, **6**.

705 Lusher, D., Koskinen, J., Robins, G., Lusher, D., Koskinen, J. & Robins, G. (2013). Exponential random
706 graph models for social networks. Structural analysis in the social sciences.

707 Marcum, C.S. and Butts, C.T. (2015). Constructing and Modifying Sequence Statistics for relevent
708 Using informR in R. Journal of Statistical Software, 64(5), 1-36. URL
709 http://www.jstatsoft.org/v64/i05/Martin, J.L. (1999). A General Permutation-Based QAP
710 Analysis Approach for Dyadic Data from Multiple Groups. *Connections*, **22**, 50–60.

711 Mizruchi, M.S. & Neuman, E.J. (2008) The effect of density on the level of bias in the network
712 autocorrelation model. Social Networks, 30, 190–200.

713 Neuman, E.J. & Mizruchi, M.S. (2010) Structure and bias in the network autocorrelation model.
714 Social Networks, 32, 290–300.

715 Nightingale, G., Boogert, N.J., Laland, K.N. & Hoppitt, W. (2014). Quantifying diffusion in social
716 networks: a Bayesian approach. *Animal social networks. Oxford University Press, Oxford*, 38–52.

717 Páez, A., Scott, D.M. & Volz, E. (2008) Weight matrices for social influence analysis: An investigation
718 of measurement errors and their effect on model identification and estimation quality. Social
719 Networks, 30, 309–317.

720 Read, J.M., Eames, K.T.D. & Edmunds, W.J. (2008). Dynamic social networks and the implications for
721 the spread of infectious disease. *Journal of The Royal Society Interface*, **5**, 1001–1007.

722 Reynolds, J.J.H., Hirsch, B.T., Gehrt, S.D. & Craft, M.E. (2015). Raccoon contact networks predict
723 seasonal susceptibility to rabies outbreaks and limitations of vaccination. *Journal of Animal*
724 *Ecology*, **84**, 1720–1731.

725 Ripley, R., Boitmanis, K. and Snijders, T.A.B. (2013). RSiena: Siena - Simulation Investigation for
726 Empirical Network Analysis. R package version 1.1-232. http://CRAN.R-
727 project.org/package=RSienaRipley, R.M., Snijders, T.A.B. & Preciado, P. (2011). Manual for
728 RSIENA. *University of Oxford, Department of Statistics, Nuffield College*, **1**.

729 Robins, G., Pattison, P., Kalish, Y. & Lusher, D. (2007). An introduction to exponential random graph

730 (p*) models for social networks. *Social networks*, **29**, 173–191.

731 Rohani, P., Zhong, X. & King, A.A. (2010). Contact network structure explains the changing
732 epidemiology of pertussis. *Science*, **330**, 982–985.

733 Shalizi, C.R. & Rinaldo A. (2013). Consistency under sampling of exponential random graph models.
734 *Annals of Statistics*, **41**, 508-535.

735 Shaw, L.B. & Schwartz, I.B. (2008). Fluctuating epidemics on adaptive networks. *Physical Review E*,
736 **77**, 66101.

737 Silk, M.J., Jackson, A.L., Croft, D.P., Colhoun, K. & Bearhop, S. (2015). The consequences of
738 unidentifiable individuals for the analysis of an animal social network. *Animal Behaviour*, **104**,
739 1–11.

740 Silk, M.J., Croft, D.P., Delahay R.J., Hodgson, D.J., Boots M., Weber N. and McDonald R.A. (2017).
741 Using Social Network Measures in Wildlife Disease Ecology, Epidemiology, and Management.
742 BioScience doi: 10.1093/biosci/biw175

743 Smith, J.A. & Moody J. (2013). Structural effects of network sampling coverage I: Nodes missing at
744 random. *Social Networks*, **35**, 652-688.

745 Smith, J.A., Moody, J. & Morgan J.H. (2017). Network sampling coverage II: The effect of non-random
746 missing data on network measurement. *Social Networks*, **48**, 78-99.

747 Snijders, T.A.B., Van de Bunt, G.G. & Steglich, C.E.G. (2010). Introduction to stochastic actor-based
748 models for network dynamics. *Social networks*, **32**, 44–60.

749 Stehlé, J., Voirin, N., Barrat, A., Cattuto, C., Colizza, V., Isella, L., Régis, C., Pinton, J.-F., Khanafer, N. &
750 Van den Broeck, W. (2011). Simulation of an SEIR infectious disease model on the dynamic
751 contact network of conference attendees. *BMC medicine*, **9**, 1.

752 Tranmer, M., Marcum, C.S., Morton, F.B., Croft, D.P. & de Kort, S.R. (2015). Using the relational
753 event model (REM) to investigate the temporal dynamics of animal social networks. *Animal*
754 *behaviour*, **101**, 99–105.

755 Tunc, I. & Shaw, L.B. (2014). Effects of community structure on epidemic spread in an adaptive
756 network. *Physical Review E*, **90**, 22801.

757 VanderWaal, K.L., Atwill, E.R., Isbell, L.A. and McCowan, B. Linking social and pathogen transmission
758 networks using microbial genetics in giraffe (*Giraffa camelopardalis*). Journal of Animal
759 Ecology, **83**¸406-414.

760 Wang, B., Cao, L., Suzuki, H. & Aihara, K. (2010). Epidemic spread in adaptive networks with
761 multitype agents. *Journal of Physics A: Mathematical and Theoretical*, **44**, 35101.

762 Webb, S.D., Keeling, M.J. & Boots, M. (2007a). Host–parasite interactions between the local and the
763 mean-field: How and when does spatial population structure matter? *Journal of Theoretical*
764 *Biology*, **249**, 140–152.

765 Webb, S.D., Keeling, M.J. & Boots, M. (2007b). Spatially extended host–parasite interactions: the role
766 of recovery and immunity. *Theoretical population biology*, **71**, 251–266.

767 Weber, N., Carter, S.P., Dall, S.R.X., Delahay, R.J., McDonald, J.L., Bearhop, S. & McDonald, R.A.
768 (2013). Badger social networks correlate with tuberculosis infection. *Current Biology*, **23**, R915–
769 R916.

Welch, D., Bansal, S. & Hunter, D.R. (2011). Statistical inference to advance network models in epidemiology. *Epidemics*, **3**, 38–45.

White, L.A., Forester, J.D. & Craft, M.E. (2015). Using contact networks to explore mechanisms of parasite transmission in wildlife. *Biological Reviews*.

Wilson, J.D., Denny, M.J., Bhamidi, S., Cranmer, S.J. & Desmarais, B.A. (2017) Stochastic weighted graphs: Flexible model specification and simulation. *Social Networks*, **49**, 37–47.

List of supplementary information

1. **Supplementary Material - The application of statistical network models in disease research:** Word document containing a description of and results from the four example analyses used in the paper, together with the annotated R code for implementing these examples.
2. **Ages.csv:** Age data for use in network autocorrelation model and exponential random graph model examples
3. **Complete Membership.csv:** Social community membership for use in network autocorrelation and exponential random graph model examples
4. **indivsexes.csv:** Sex data for use in network autocorrelation model and exponential random graph model examples
5. **overallnetwork.csv:** Network data for use in network autocorrelation model and exponential random graph model examples
6. **TBstatsF.csv:** bTB infection data for use in network autocorrelation model and exponential random graph model examples
7. **autumnmatrix.csv:** binary autumn network for use in stochastic actor-oriented model and temporal exponential random graph model examples
8. **summermatrix.csv:** binary summer network for use in stochastic actor-oriented model and temporal exponential random graph model examples
9. **wintermatrix.csv:** binary winter network for use in stochastic actor-oriented model and temporal exponential random graph model examples
10. **grouplocsSAOM.csv:** group location data for use in the stochastic actor-oriented model example
11. **MembershipSAOM.csv**: Social community membership data for use in the stochastic actor-oriented model example
12. **SAOMsexes.csv**: Sex data for use in the stochastic actor-oriented model example
13. **SAOMTBstats.csv**: bTB infection data for use in the stochastic actor-oriented model example
14. **MembershipTERGM.csv**: Social community membership data for use in the stochastic actor-oriented model example
15. **TERGMsexes.csv**: Sex data for use in the stochastic actor-oriented model example
16. **TERGMTBstats.csv**: bTB infection data for use in the stochastic actor-oriented model example

Figures and Tables

Table 1. The advantages and disadvantages of the main statistical modelling approaches to studying contact networks for disease.

Model	Dependent variable	Network type	When to use	Advantages	Disadvantages	Mathematical details	Software
Generalised linear (mixed) model (*GLM/GLMM*)	Individual traits	Static/ Dynamic	Can be used to test a whole range of hypotheses related to network position (with appropriate randomisations) *E.g. Do network positions of individuals infected with PathogenX show distinct properties from those of uninfected individuals?*	-Familiarity of researchers -Well-developed methods in animal social networks -Can be used with group-based or association-based methods of network construction more easily	-Not specifically designed to incorporate non-independence implicit to networks -System specific randomisations required that generate uncertainty around the null hypothesis rather than the observed parameter	Croft *et al.* (2011) Farine and Whitehead (2015)	*lme4(modelling) igraph/asnipe (randomisations)*
Temporal network autocorrelation model (*TNAM*)	Individual traits	Static/ Dynamic	For testing hypotheses about how individual traits change in the context of a network in a single network or series of network snapshots. *E.g. How do network position, past network position and the infection status of neighbouring individuals best explain infection with pathogenX?*	-Can be used to explicitly account for non-independence of network data -Enables the direct and indirect effects of other individuals in the network to be modelled. -Same modelling framework can be applied to static and dynamic (multiple network snapshots) networks	-Network dependency must be defined by user and goodness of fit cannot be tested - Complex to include interactions between more than two variables. [It is possible if the model matrix is generated using the function tnamdata()] -Robustness when used in group-based or association-based networks or with randomisation-based hypothesis testing unknown.	Doreian *et al.* (1989) Leenders (2002) Hays *et al.* (2010)	*xergm (tnam)*
Multiple regression quadratic assignment procedure (*MRQAP*)	Edge values	Static	For testing hypotheses about how relational traits are affected by other dyadic variables (i.e. matrix correlations) *E.g. Does there tend to be a difference in interaction strength between susceptible-susceotible and susceptible-infected dyads*	-Familiar to ecologists -Accessible method to implement -Can be used to analyse association-based animal networks	-No opportunity to model dependency structure of network -No standard errors estimated around model parameters -Problems in sparse networks and with collinear explanatory variables	Martin (1999) Dekker, Krackhardt & Snijders. (2007)	*sna, asnipe*
Exponential random graph model (*ERGM*)	Network topology	Static	For testing hypotheses about the properties of edges or local network topology in a single network. *E.g. How does pathogenX infection affect an individual's social relationships?*	-Modelling framework accounts for conditional dependence within the network -Models the edges themselves, which are often of most interest from an epidemiological pespective -Can include structural effects of biological interest or control for study design/social system e.g. distance, group membership	-Lack of flexibility to have interaction terms within the model *Nb. It is possible to set up use defined terms but this is will be challenging* -Restricted to interaction- or contact-based network in which the researcher is confident of ties (use for group-based networks untested)	Robins *et al.* (2007) Lusher *et al.* (2013)	*ergm, ergm.count, GERGM*
Latent space model	Edge values	Static/ Dynamic	For testing hypotheses about the properties of dyads in a single network (no inclusion of network topology). *E.g. How does pathogenX infection affect an individual's social relationships?*	-Modelling framework accounts for conditional dependence within the network -Models the edges themselves, which are often of most interest from an epidemiological pespective -Generally simpler implementation and fitting than ERGMs as dependencies estimated automatically	-Hypotheses related to network topology cannot be tested as network dependencies are included in the latent space component -Lack of flexibility to have interaction terms within the model -Use in association-based networks untested -Interpretation of coefficients can be complex if correlated with the positions of nodes in latent space -User-defined definitions of latent space need to completed with caution	Hoff, Raftery, & Handcock (2002) Krivitsky *et al.* (2009)	*latentnet*

Network-based diffusion analysis (*NBDA*)	Transmission process	Static	For testing the hypothesis that the acquisition of trait on a static network is a social process. *E.g. Does the spread of pathogenX depend on contact network structure*	-Simple to implement with a clear hypothesis test (whether the acquisition of a trait is best explained by social or non-social processes) that is highly relevant to disease research	-Lack of flexibility -Only takes into account a single static network structure (cf. tnam)	Franz and Nunn (2009) Nightingale *et al.* (2014)	*code available online (see main text)* *spatialnbda*
Stochastic actor-oriented model (*SAOM*)	Network topology and individual traits	Dynamic	For testing hypotheses related to how a trait influences an individual's dynamic network position or for testing hypotheses about how a trait and an individual's social network position are inter-related *E.g. How does infection with PathogenX covary with social behaviour?*	-Accounts for conditional dependence within the network -Can model both the probability of edges over time and differences in rates of network change depending on structural effects, and nodal and dyadic covariates	-Restricted to interaction- or contact-based network in which the researcher is confident of ties. Use for association-based or group-based networks untested. -Only possible to use for binary or ordinal networks -Excessive changes in network composition over time can lead to estimation problems	Snijders *et al.* (2010)	*Rsiena*
Temporal exponential random graph model (*TERGM*)	Network topology	Dynamic	For testing hypotheses about the properties of edges or local network topology in a series of network snapshots. *E.g. How stable are social relationships and how does infection with PathogenX affect this?*	Modelling framework accounts for conditional dependence within the network -Models the edges themselves, which are often of most interest from an epidemiological pespective -Can include structural effects of biological interest or control for study design/social system e.g. distance, group membership - Temporal covariates enable tests of interaction stability and can interact with covariates to test how this affected by dyadic covariates -Able to provide user-defined functions (which can be non-linear)for temporal change in network structure	-Lack of flexibility to have interaction terms within the model *Nb. It is possible to set up use defined terms but this is will be challenging* -Restricted to interaction- or contact-based network in which the researcher is confident of ties (use for group-based networks untested) -Relative to SAOMs, less informative about rates of network change over time -Missing data has to be imputed or the individuals removed from the network	Hanneke *et al.* (2010) Leifeld *et al.* (2014)	*xergm (btergm)*
Relational event model (*REM*)	Interaction or contact events	Temporally -explicit Dynamic	For testing hypotheses about the timing and patterns of interactions or contacts in temporally-explicit data. *E.g. Is the temporal pattern of social contacts different for individuals infected with PathogenX?*	-Temporally-explicit -Support constraints make the framework very adaptive as to appropriate datasets -Does not require individuals to be present for the entire study period	-More complex implementation and interpretation -Harder to test hypotheses directly related to network structure and position than other approaches; this often has intuitive appeal for disease research. - Computationally intensive for larger networks and/or more complex models as a result of maintaining temporally-explicit data.	Butts (2008)	*relevent (+informR), rem*

Table 2. Details of the type of model term, what type of network to use it in and guidance on how and when to use it for a selection of standard terms to consider when using ERGMs and TERGMs.

ERGM term	Network type	Term type	Use to…
edges *density*	Binary	Structural	Similar to an intercept in a GLM - gives the probability of edges in the network relative to a random network. Density is equivalent to edges divided by n(n-1)/2
non-zero	Weighted	Structural	Zero-inflation term in weighted networks (accounts for the fact that most networks are sparse and therefore distribution of edge weights is zero-inflated)
sum	Weighted	Structural	Similar to the intercept in a GLM for weighted networks
kstar(x:y)	Binary	Structural	A statistic for each kstar between x and y . kstar(1) is equivalent to edges
triangle *localtriangle(x)*	Binary	Structural	A statistic for the number of triangles in the network (i.e. a measuring of clustering/transitivity). localtriangle(x) calculates only triangles between neighbours which are given using an indicator matrix x.
transitiveweights *cyclicalweights*	Weighted	Structural	Both of these terms can be used to calculate triangles in weighted networks taking into account the weights of edges
nodefactor(x)	Both	Node-based	The effect of a categorical nodal variable on the probability/weight of edges
nodecov(x)	Both	Node-based	The effect of a continuous nodal variable on the probability/weight of edges
nodematch(x)	Both	Node-based	The probability/weight of edges between two individuals of the same versus different values of a categorical nodal variable. The argument diff=TRUE can provide separate estimates for each level of the factor
absdiff(x)/ *absdiffcat(x)*	Both	Node-based	The effect of the difference in values of a continuous nodal variable between nodes on the probability/weight of an edge formed between them.
edgecov(x)/dyadcov(x)	Both	Dyad-based	The effect of a dyadic covariate (e.g. relatedness) on the probability/weight of edges formed. Using dyadcov(x) applies directed covariates when the network itself is directed
memory(type="")	Both	Temporal	The stability of edges over time. Additional arguments in type can be used to test different memory effects e.g. all potential edges ("stability") or only complete edges ("autoregression")
timecov(x,transform= *function(t))*	Both	Temporal	Trends in edge formation over time (nature of trend given by transform argument). Can additionally include a dyadic covariate x to create an interaction effect

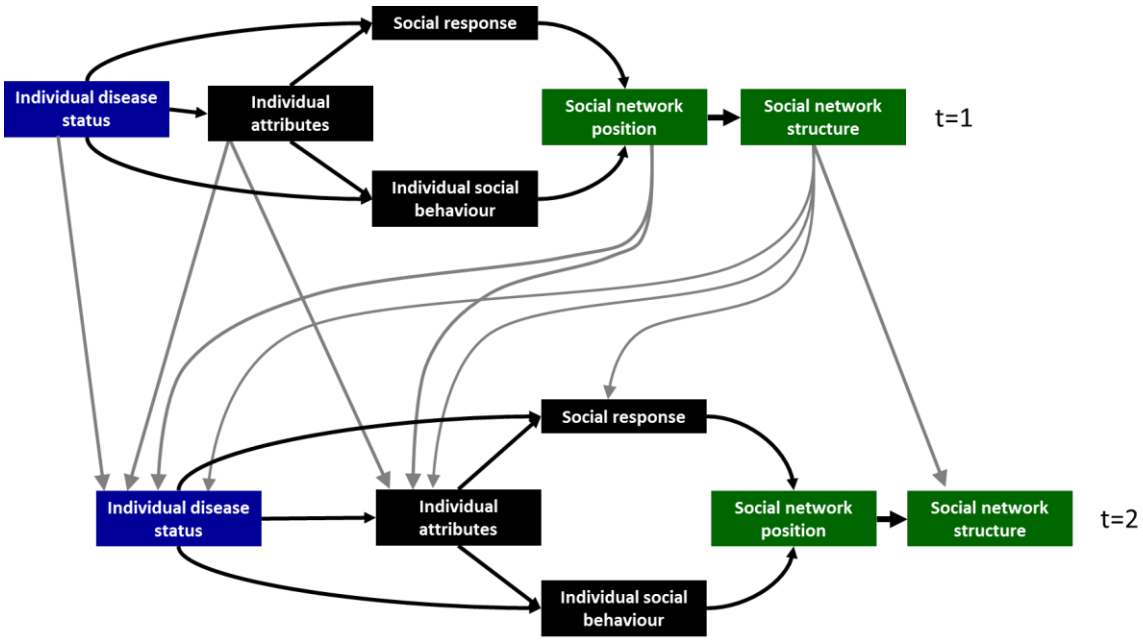
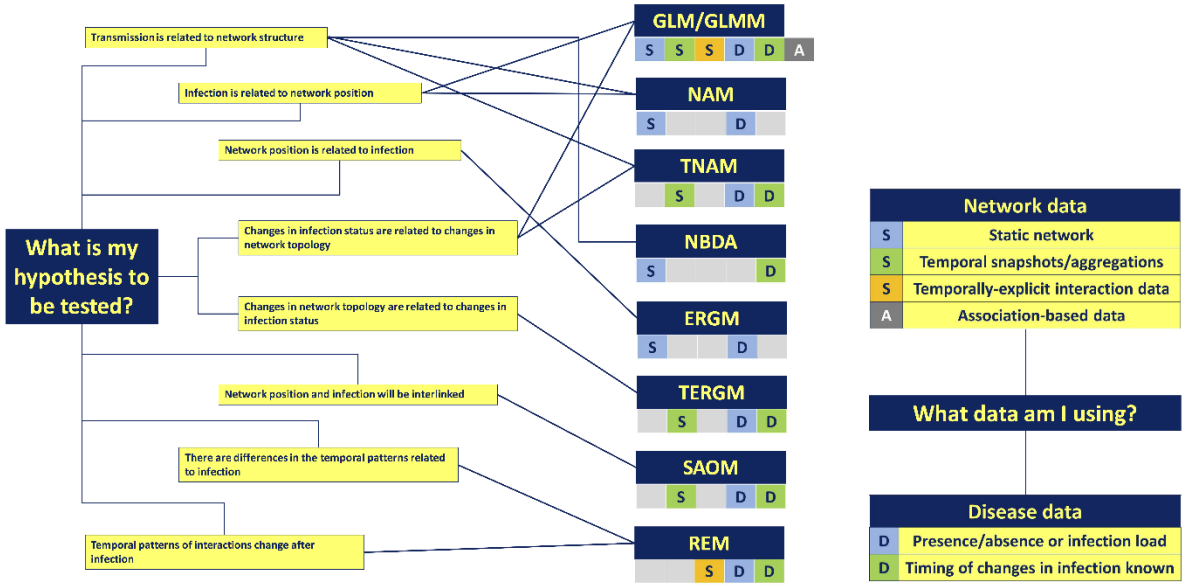


Figure 1. The dynamics of social interactions and disease across two time points (t=1 and t=2). Models of static networks can only explore correlations at one point in time; by incorporating dynamic modelling approaches it is possible to explore causation. Individual attributes in this graph refer to both fixed phenotypic traits such as sex, and conditional traits such as physiological stress, immunocompetence and condition. Social response represents the social behaviour of other individuals towards a focal individual.



831

832 Figure 2. A guide to statistical model use to test hypotheses about the relationship between social
833 contacts/interactions and disease for the most appropriate models to test hypotheses about
834 networks and disease. GLM is generalised linear model, GLMM is generalised linear mixed model,
835 ERGM is exponential random graph model, NBDA is network-based diffusion analysis, SAOM is
836 stochastic actor-oriented model, TERGM is temporal exponential random graph model and REM is
837 relational events model.