Generation and applications of simulated datasets to integrate social network and demographic analyses

Matthew J Silk[1]* and Olivier Gimenez[1]

1 CEFE, Univ Montpellier, CNRS, EPHE, IRD, Montpellier, France

*corresponding author: matthewsilk@outlook.com

Running headline: Tools for integrated network-demographic models

Abstract

1. Social networks are tied to population dynamics; interactions are driven by population density and demographic structure, while social relationships can be key determinants of survival and reproductive success. However, difficulties integrating models used in demography and network analysis have limited research at this interface.
2. Here we introduce the R package genNetDem for simulating integrated network-demographic datasets. By generating co-capture data with known statistical relationships it provides functionality for methodological research. We demonstrate its use with case studies testing how imputation and sampling design influence the success of adding network traits to conventional Cormack-Jolly-Seber (CJS) models.
3. We show that incorporating social network effects in CJS models generates qualitatively accurate results, but with downward-biased parameter estimates when network position influences survival. Biases are greater when fewer interactions are sampled or fewer individuals observed in each interaction. However, model performance is robust to network covariance, network structure and sampling design. Longitudinal imputation of missing network measures is more stable than cross-sectional imputation but neither approach successfully generates accurate parameter estimates.
4. While our results indicate the potential of incorporating social effects within demographic models, they show that imputing missing network measures alone is insufficient to accurately estimate social effects on survival. This points to the importance of incorporating network imputation approaches. genNetDem provides a flexible tool to aid these methodological advancements and help researchers testing other sampling considerations in social network studies.

Key-words: co-capture data; hidden Markov model; population dynamics; stochastic block model; survival;

Introduction

Network analysis has revolutionised animal social behaviour research by quantifying how dyadic social interactions and relationships are nested in wider group- and population-level social structures (J. Krause et al., 2014; Pinter-Wollman et al., 2013). Network studies in behavioural ecology have often focussed on how the position of an individual within its social network influences its fitness, either via reproductive success (Formica et al., 2012; Oh & Badyaev, 2010) or survival (Blumstein et al., 2018; Ellis et al., 2017; Stanton & Mann, 2012).

Quantifying direct links between social network position and fitness can help us understand how selection acts on social behavioural traits. Further, determining how social behaviour is linked to survival can identify demographic consequences of interactions and associations (Clements et al., 2022), which can help develop more realistic models for how social species respond to population declines or environmental change (Snijders et al., 2017). However, while there is growing interest in linking animal social networks with demography (Shizuka & Johnson, 2020), there remain many methodological challenges.

Currently most studies that link network position and fitness use known fate approaches such as generalised linear models (e.g. (Blumstein et al., 2018)) or Cox proportional-hazards models (e.g. (Ellis et al., 2017)). However, application of these approaches is limited in many wild populations where individuals that are alive are not necessarily detected. In these cases survival is most commonly estimated using hidden Markov models (HMMs; (McClintock et al., 2020)) that can simultaneously estimate survival and capture probabilities (Gimenez et al., 2012; Pradel, 2005). These models also have potential as tools in animal social network analysis (Clements et al., 2022; Fisher et al., 2017), especially when not all associations are detected. However, it is challenging to provide universal guidance on the applicability of these approaches given the diversity of animal social systems and sampling designs used to study them.

Here we introduce the R package genNetDem to simulate co-capture datasets. We define a co-capture dataset as one in which a capture-recapture data also provides information on social structure, such as when individuals are caught or observed in groups (see also (Silk et al., 2021)). The package generates integrated longitudinal social network and capture-recapture datasets with known statistical relationships. This provides functionality for methodological research, power analyses and sampling design. Here we introduce the package, describe key functions and outline effective workflows. We then provide two case studies to demonstrate its use. Finally, we identify key next steps in merging social network and demographic analyses, and discuss the role of genNetDem in these.

genNetDem overview

genNetDem is a set of R (R Core Team, 2021) functions that generate longitudinal social network and/or capture-recapture datasets with known underlying properties. Functionality can be split into four broad groups: a) population features; b) survival features; c) social network features; and d) observation features. The package is modular meaning specific components can be used in isolation or user-generated code can be integrated to extend functionality to different ecological or social contexts.

genNetDem is available on GitHub (https://github.com/matthewsilk/NETDEM/tree/main/genNetDem). We provide a brief overview of key functions in the main text and more detail in the Supplementary Materials.

Population features

The population features provide capability to simulate a population and generate data about individuals in it. There is then functionality that simulates population dynamics based on individual survival probabilities (see *survival features*) and stochastic recruitment that maintains an approximately stable population size when employed.

The **population_generation_basic()** function generates data for a group-structured population distributed uniformly in 2D space. The function takes two arguments: *n* defines the population size and *ng* the number of groups in the population. When *n*=*ng* individuals are distributed uniformly across the defined coordinates. When *n*>*ng* groups are distributed uniformly across the same coordinates with individuals in the same group sharing the same spatial location. Currently, simulated population size is independent of the extent of the area it occupies. Therefore, population density will increase with population size and impact spatial effects on social network structure. This does not represent a problem except when users want to compare the social structures of populations of different sizes. The **indiv_info_gen()** and **indiv_info_add()** functions provide flexibility in generating and updating individual-level trait data. Variables can be specified as covariates (e.g. size) or categorical factors (e.g. sex), with further arguments specifying additional features of the variable (e.g. the distribution of a covariate or the number of levels and level names of a factor).

The **timestep_demographics()** function controls survival and recruitment in the simulated population. Survival is stochastic based on each individual's survival probability (see *survival features*). The number of recruits is Poisson distributed (λ calculated from survival probability). When the population is group-structured individuals can only be recruited into existing groups.

Survival features

The **covariates_survival()** function allows survival probabilities to be calculated for each individual based on individual traits or the position of an individual within a population social network (this could be any network provided to the function; the underlying social networks, simulated interaction network or a separate user-specified network). There is considerable flexibility in which measures of network position can be included as covariates; both the function and R package used can be specified within the function, with functionality for most common packages (e.g. sna: (Butts, 2014); igraph: (Csardi & Nepusz, 2006); tnet: (Opsahl, 2009)) incorporated. It is also possible to simulate network covariance in survival whereby closely connected individuals have either more or less similar survival probabilities than expected by chance. Currently, covariates_survival() simulates independent (additive) effects of traits, meaning there is no functionality to capture interactions among variables (e.g. network position having different effects in males than females). The simpler **basic_survival()** function generates population-level survival probabilities in the absence of covariates.

Network features

There are two core functionalities of the network features: to generate underlying social networks for the population; and to generate grouping events (interactions/associations) based on these networks. There are also two ***network_checker()*** functions that quantify and visualise how well social networks derived from grouping events match the underlying network used to generate them.

The **network_generation_covariates()** function generates an underlying network structure based on group membership (as defined when generating the population), spatial locations and individual traits. Figure 1 shows examples of networks generated. Current functionality is focussed on how these traits impact the probability of forming social connections within and between groups separately, thus employing a stochastic block model (Lee & Wilkinson, 2019). Therefore, genNetDem currently lacks functionality to incorporate some known social processes such as triadic closure or assortativity (e.g. females being more closely connected to other females). It is also currently not possible for interaction effects to be coded directly (e.g. if size effects on connectivity were different for males and females). Edge probabilities and edge weights are modelled independently to allow variables to explain variation in one or both of them. There is a simpler **network_generation_basic()** function that uses the same generative model without covariates.

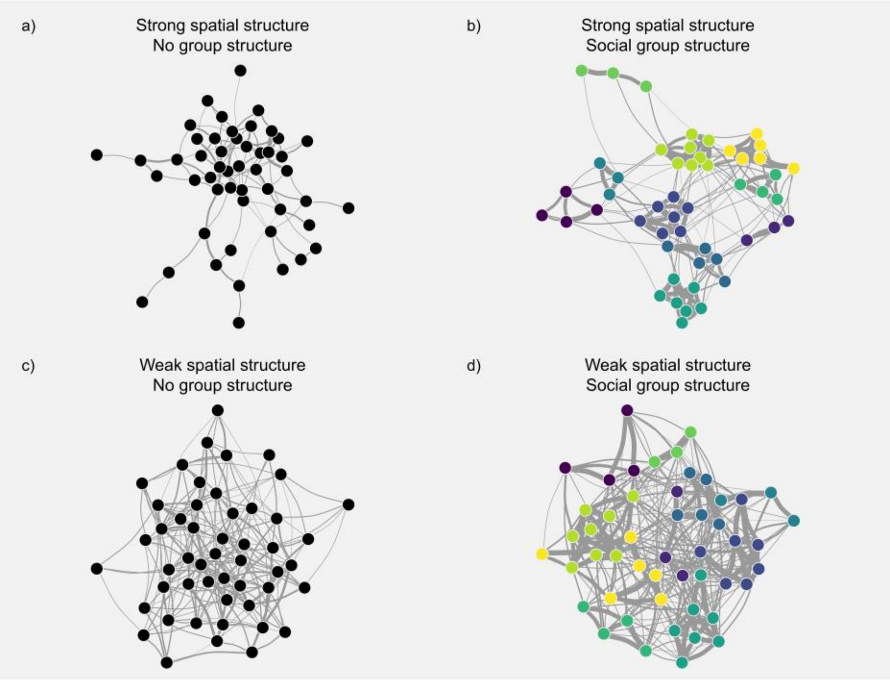


Figure 1. Examples of the diverse underlying social network structures it is possible to simulate with genNetdem. Here we explore the impact of spatial and social structure on otherwise similar sets of rules for the generation of social relationships. Code to replicate this figure and further explore network possibilities is provided in the Supplementary Materials.

When modelling longitudinal network data, the individual social network positions could vary from relatively stable to highly dynamic (Pinter-Wollman et al., 2013). The **network_rewire_covariates()** function adds newly recruited and removes dead individuals from the network but also provides functionality to select probabilities that a) an individual changes some social relationships and b) each social relationship for selected individuals changes. This allows flexibility in how dynamic simulated networks are.

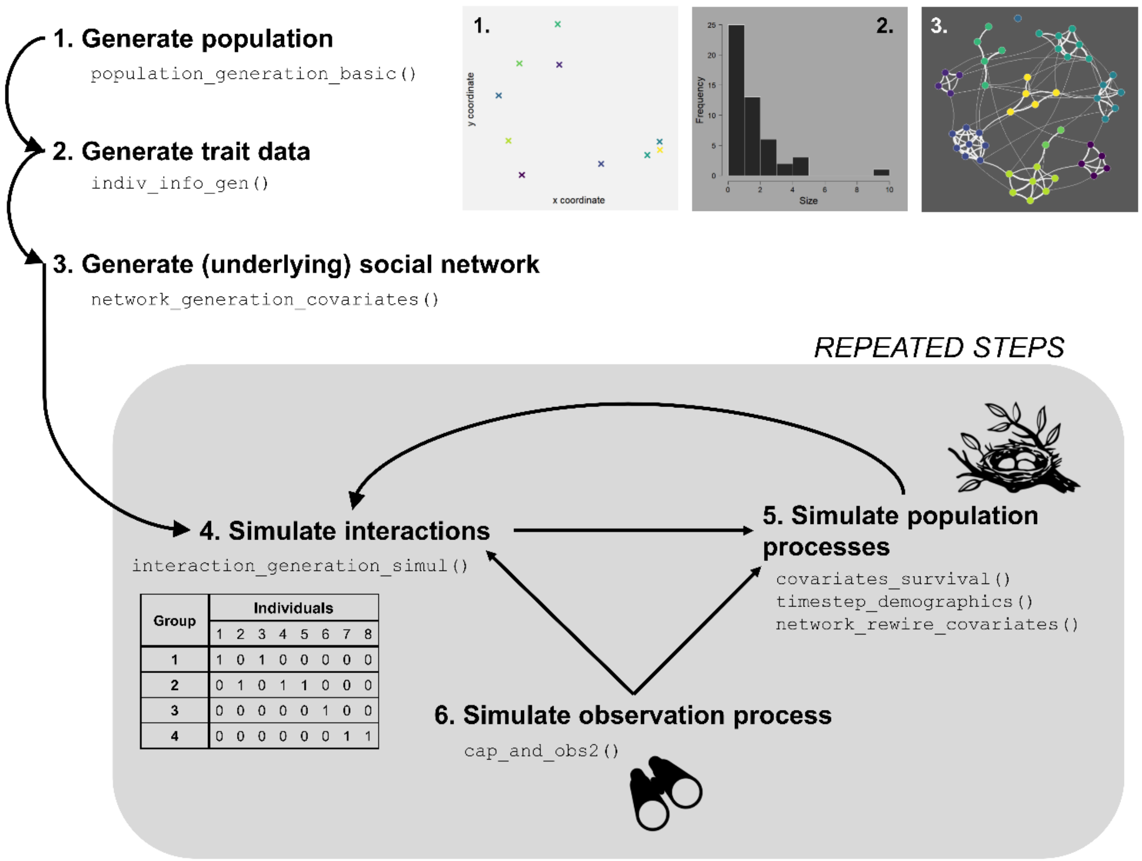
There are two functions that generate grouping events based on underlying network structure: **interaction_generation_simul()** and **interaction_generation_seq()**. The difference between them is that the former divides all individuals in the population into groups (or isolates) at each time point, while the latter independently samples one group of a defined size from the population at a time. The former is more widely useful. It uses data on individual IDs, their underlying social network and a mean group size to divide the populations into groups, with group membership being stored in a group-by-individual matrix (GBI; see (Farine, 2013)). The n_ts argument defines the number of times this process is repeated (i.e. number of "***behavioural timesteps***"). Assigning individuals into groups based on the underlying network can create computational challenges if unconstrained. We use a similar approach to (Evans et al., 2020), with individuals being added to groups sequentially and the probability of joining being proportional to the strength of its social relationships with existing members (see Supplementary Materials). Including a non-zero float argument means it is never impossible to add an individual to an existing group even in the absence of any social connections. While it may be tempting to reduce the float to zero this can result in it being impossible (or computationally challenging) to successfully sample all individuals into groups. However, care should be taken with particular combinations of group size distributions and underlying network structures that these relaxations do not dominate interaction generation. This can be checked with the **network_checker_simul()** function.

Observation features

The observation features sample the simulated grouping events and generate data for subsequent analyses. Data is generated in a variety of formats including GBI matrices for social network analysis and classic capture-recapture formats. There are two **cap_and_obs()** functions that generate an observed network dataset based on the sampling strategy and design. Of the two **cap_and_obs2()** has greater flexibility (see Supplementary Materials). Inputs include: a) data on true grouping events (the GBI and a vector indicating which behavioural timestep each group occurred in); b) vectors indicating behavioural timesteps to be sampled, indicated separately for captures and observations; c) the success of sampling including both the proportion of groups detected and the proportion of individuals in each sampled group detected; and d) a vector indicating which (if any) individuals had been captured previously. The function returns GBIs for captured and observed groups and other related information. The **cap_dat_gen()** function transforms these network datasets into capture histories for both behavioural timesteps and demographic timesteps.

Workflows

While genNetDem is designed to be modular so that individual components can be adjusted to perform a range of tasks, many of the functions fit well within specific workflows. We illustrate one such common workflow (Fig. 2), but various other applications are demonstrated package vignettes.



Group	Individuals							
	1	2	3	4	5	6	7	8
1	1	0	1	0	0	0	0	0
2	0	1	0	1	1	0	0	0
3	0	0	0	0	0	1	0	0
4	0	0	0	0	0	0	1	1

Figure 2. An example workflow for using genNetDem to simulate integrated network-demographic datasets. This is a simplified version of the approach used in the case studies with the grey box capturing a demographic timestep and step 4 involving one or more behavioural timesteps. Note that multiple set of interactions can be generated prior to the simulation of population processes if desired. The modular nature of the functions in the package mean that different parts of this workflow can also be used independently. Further usage examples are provided in the package vignettes.

Case Studies

We use two complementary case studies to illustrate the use of genNetDem. In the first we test how our ability to estimate the relationship between network position and survival depends on sampling effort; whether local or global centrality affects survival; and network dynamics. We compare the performance of cross-sectional versus longitudinal imputation of the network position of non-detected individuals and explore the importance of network covariance in survival probabilities. In the second we demonstrate how a researcher could use genNetDem to compare sampling designs. We test how the power to estimate relationship between network position and survival depends on how sampling effort is distributed through time. Our simulation asks the question as to whether it is better to concentrate resources into intensively monitoring more groups in fewer sampling windows or fewer groups in more sampling windows. We examine whether any differences are impacted by the proportion of individuals detected in each sampled group and the structure of the underlying social network.

Methods common to both case studies

In both case studies we use genNetDem to simulate survival and social interactions and then sample from them to generate capture histories. We fit hidden Markov Models to estimate survival and capture probabilities using nimble (de Valpine et al., 2017, 2022).

Data recorded from simulation runs

We recorded a) the capture-recapture dataset for each demographic timestep; b) the sampled social network generated from all observed interactions within each demographic timestep; c) individual survival probabilities for each demographic timestep; and d) information on true population size and the number of individuals recorded at each demographic timestep. We estimated the network measure of interest from the sampled social network and scaled it within each demographic timestep to use as an explanatory variable.

Modelling approach

We fitted Cormack-Jolly-Seber models estimating both capture and survival probabilities (Lebreton et al., 1992) and used Bayesian inference for parameter estimation. We included explanatory variables of sex and social network measure (either strength or betweenness). In each model we used weakly informative priors for all parameters (Gaussian distribution with $\mu=10$ and $\sigma=10$ for survival-related variables, uniform distribution between 0 and 1 for capture probability). We used a single Markov chain of 3000 iterations with a burn-in of 500 and a thinning interval of 5. We confirmed that this number of iterations was typically sufficient for model convergence and an adequate effective sample size in a subset of simulations.

Analysis of simulation results

From each simulation run we calculated the posterior median and standard deviation, the proportion of the posterior greater than zero, and the 89% HDI. We also calculated a binary variable indicating whether or not 0 was contained within the 89% HDI. We could then compare model performance visually and by calculating statistical clarity for positive social effects on survival as the proportion of simulation runs where 0 fell outside the 89% HDI.

Case study 1: Performance of basic imputation to estimate social effects on survival

Specific Methods

Overview of Data generation

We simulated a population of 200 individuals with no underlying group structure. Individual variation was restricted to a single two-level categorical variable – sex. The underlying social network had moderate spatial structure.

We simulated the behaviour and survival of individuals over 10 demographic timesteps (over which survival was simulated), each containing five behavioural timesteps (at which individuals were organised into grouping events). Grouping events had a mean size of two individuals (many events were dyadic and individuals were frequently alone) to capture a situation where a species rarely occurs in large aggregations. Survival probability depended on sex (moderate effect of 0.5 on a logit scale) and position in the social network calculated from grouping events (see below) with a baseline survival probability of 0.8 in females. We assumed no recruitment into the population (i.e. the population declined over the simulation).

We assumed that all individual in the population were marked or individually-identifiable prior to the start of the study. Captures and/or observations (which were functionally equivalent as all individuals were marked) took place in all behavioural timesteps (50 in total). Each group had either a 25%, 50% or 75% percent chance of being detected (parameter varied between simulation runs) with the detection probability of an individual in a detected group fixed at 0.9.

Simulation structure

In total we generated 3240 simulated datasets, varying five parameters that influenced network dynamics (one parameter), network effects on survival (three parameters) and sampling (one parameter).

- Network dynamics: we varied the probability that an individual’s existing connections in the underlying social network were rewired with values set at 0 (no rewiring), 0.1 and 0.5. The per-edge probability that an individual changed its connections was 0.5.
- Network effects on survival: a) we varied the network measure that influenced survival to be either strength (local measure; sum of weighted connections) or betweenness (global measure; number of shortest paths passing through an individual); b) we varied

the effect size to be 0 (no effect), 0.4 (moderate effect) or 0.8 (strong effect); c) we altered covariance of individual survival within the network to be negative (strongly connected individuals have more dissimilar survival probabilities), neutral or positive (strongly connected individuals have more similar survival probabilities).

- Sampling: we varied the probability of sampling (either capturing or observing) a group at each behavioural timestep to be 0.25, 0.5 or 0.75.

For each combination of parameters (162) we ran 20 replicates.

Data recorded

In addition to the four types of data described in the combined methods, we also recorded the full social network generated from all interactions within each demographic timestep (including those not observed). We estimated the network measure of interest from these full networks and scaled them within each demographic timestep as for measures from partial networks.

Model-fitting

From each simulation run we fitted four model versions (see combined methods for details on model-fitting). The four versions differed in: a) using the measure from the sampled network and a longitudinal approach for imputing non-observed individuals; b) using the measure from the sampled network and cross-sectional imputation; c) using the measure from the full (unobserved) network and longitudinal imputation; and d) using the measure from the full network and cross-sectional imputation. For cross-sectional imputation missing values were estimated using the mean and variance of the (scaled) focal network measure from a given demographic timestep. For longitudinal imputation missing values were estimated using the mean and variance of the focal network measure for each individual where possible and the overall mean and variance when not.

Analysis of simulation results

Prior to the general analysis outlined above, we assessed whether the model had converged using the posterior median and standard deviation of its estimate for the social effect on survival. We used k means clustering to identify groups of simulation runs where the model was unlikely to have converged. We used k=6 clusters and retained 3 out of 6 of these clusters based on the elbow method and visual inspection of the output (Fig. S1). This method identified ~2.5% of models had likely not converged.

To compare the success of models that used network measures calculated from the partial versus full network we calculated the earth mover's distance (EMD) of the posterior distributions (Touzalin et al., 2022) for the parameter of interest from relevant pairs of models. (i.e. we calculated the EMD for model versions using the full and partial network together with longitudinal imputation and also the EMD for the model versions using the full and partial network together with cross-sectional imputation). EMDs provide a measure of overlap of the posterior distributions.

Results and Discussion

Overall, we show it is possible to estimate social effects survival from partial networks, albeit with substantial limitations in power (Fig. 3, Table 1, Table S1). Estimates of social effects on survival were downward biased meaning that statistical power was limited and only stronger social effects on survival are likely to be detected. Sampling effort was particularly important and interacted with how imputation was conducted in determining how well models converged and biases in parameter estimates when they did. Estimates of other parameters were unaffected.

Previous research has demonstrated that network measures from sampled, partial networks are correlated with those in the full, unobserved network but that these correlations vary depending on the proportion sampled and network measure calculated (Silk et al., 2015; Smith & Moody, 2013). Further, the regression slope is rarely 1:1 indicating values for measures estimated are not perfectly accurate (Silk et al., 2015). This likely explains many of our results showing the difficulty of detecting social effects on survival in the absence of network imputation or the use of measures from independently (and better) sampled social networks).

Network variable and covariance structure

When we compared models that used network measures from the full and partial networks we found downward-biased parameter estimates and reduced statistical clarity of results when partial network measures were used (Table 1, Fig. 3). These patterns were more striking when survival was related to a global measure of centrality (betweenness) than a local measure of centrality (strength). We found that including positive or negative covariance in survival probabilities related to social network structure had little effect on estimation or power in the contexts simulated (Fig. S3, Tables S5-6).

These results fit well within the literature on how missing individuals impact the conclusions of social network analysis, with previous studies showing that global estimates of social centrality (such as betweenness) from partial networks are less well correlated than measures of local centrality (such as strength) with equivalent measures from the full network (Silk et al., 2015). While for strength in particular downward-biased parameter estimates in combination with maintained statistical power could also be related to measures of strength being lower in the smaller, sampled network (Silk et al., 2015), this should be controlled for by scaling network measures before using them in the model. The lack of a clear effect of network covariance is somewhat surprising. These results are promising in suggesting that this may present a more limited issue in this context than often considered (e.g. (Croft et al., 2011; Farine & Carter, 2020; Silk et al., 2017)). However, the importance of covariance likely depends substantially on network structure and density, so it would be unwise to generalise these patterns without further work focussed specifically on this question.

Sampling effort, imputation approach and network dynamics

Lower sampling effort was typically associated with both a) reduced likelihood of model convergence (Table 2, Table S2), and b) downward-biased parameter estimates (Fig. 3). However, the nature of these relationships depended on the imputation approach selected (Table 1, Fig. 3), with the performance of different imputation approaches largely independent of network dynamics (Tables S3-4, Fig. S2).

Models were much less likely to converge when sampling effort was low (25% group capture probability), betweenness centrality from partial networks was used as an explanatory variable and cross-sectional imputation was used to infer missing values (Table 2, Table S2). Even when 50% of groups were sampled in these situations there was still a reduction in convergence rate. Note that this was apparent regardless of whether betweenness centrality had a positive or no effect on survival probability. Any other changes in the likelihood of model convergence were of much smaller magnitude, but generally occurred when sampling effort was low (and measures from partial networks were used).

With cross-sectional imputation and use of measures from the full network, estimation of social effects on survival were largely independent of sampling effort in the contexts examined. With longitudinal imputation there was some reduction in estimates of the social effect on survival with low levels of sampling (25% groups sampled). However, both cross-sectional and longitudinal imputation demonstrated similar relationships between sampling effort and statistical power (Table 1, Table S1), indicating that posterior distributions had higher variance when cross-sectional imputation was used.

When measures from the partial network were used instead, there was a much more substantial reduction in both parameter estimates and statistical power apparent even for higher sampling efforts (Fig. 3, Table 1). Reductions in parameter estimates were more substantial and remained linear when longitudinal imputation was used, instead flattening out for cross-sectional imputation so that the difference between 25% and 50% of groups being sampled was less than the difference between 50% and 75% (Fig. 3). However, similarly to the pattern for full network measures, this was not reflected in changes to statistical power which were broadly equivalent for both, indicating a less precise posterior distribution for cross-sectional imputation. These differences between cross-sectional and longitudinal imputation changed how EMDs calculated for the differences between posteriors from the full network and partial network model fits depended on sampling effort (Fig. 4). For cross-sectional imputation EMDs were highest for low sampling effort (p=0.25) while for longitudinal imputation they peaked at intermediate sampling effort (p=0.5). However, in general EMDs were higher for cross-sectional than longitudinal imputation.

Our results show that when social networks are constructed based on the same co-capture data used to estimate survival, even relatively small drops in sampling effort can lead to downward biases in parameter estimates and statistical power. While this pattern was especially strong when global measures of centrality such as betweenness explain variation in survival probability as expected from previous literature (Silk et al., 2015; Smith & Moody, 2013), it was also apparent when strength was associated with survival instead. However, in this latter case underestimated social effects on survival only caused substantial reductions in statistical power with very low sampling effort. Consequently, our results fit broadly within the existing literature where low sampling effort has a greater impact on global measures of centrality but suggest that missing a high proportion of interaction events leads to wider problems with subsequent statistical analyses. This was particularly apparent when cross-sectional imputation was used to estimate missing values for betweenness centrality when there was a substantial drop-off in how likely models were to converge. Combined with cross-sectional imputation generating less precise posteriors, this suggests that longitudinal imputation is a more stable option of the two, although it does lead to greater downward bias in estimates of social effects on survival. However, neither imputation approach performed well, highlighting the value of extending network imputation approaches (R. W. Krause et al., 2018, 2020; Young et al., 2020) within capture-recapture models. A good example is provided by (Clements et al., 2022), who estimate not only the network itself but also the underlying behaviours that generate the network structure within a Cormack-Jolly-Seber model. While this was done in the context of a simulation study, and so involved fitting the data-generating model, it does show the potential of network imputation to improve the accuracy of estimates of social effects on survival.

Estimates of other parameters

Estimates for other parameter values were unaffected by social effects on survival, use of measures from full or partial networks or imputation strategy (Figs. S4-6).

Table 1. Proportion of simulation runs where 0 falls outside the 89% HDI for different parameter combinations. M1: partial network - cross-sectional imputation; M2: partial network – longitudinal imputation; M3: full network – cross-sectional imputation; M4: full network – longitudinal imputation.

Network measure	**True effect**	**Model**	**Group capture probability**	**Detection rate**
Strength	0.4	M1	0.50	0.75
Strength	0.4	M2	0.50	0.72
Strength	0.4	M3	0.50	0.98
Strength	0.4	M4	0.50	0.98
Betweenness	0.4	M1	0.50	0.26
Betweenness	0.4	M2	0.50	0.25
Betweenness	0.4	M3	0.50	0.94
Betweenness	0.4	M4	0.50	0.94
Strength	0.8	M1	0.50	0.99
Strength	0.8	M2	0.50	0.99
Strength	0.8	M3	0.50	1.00
Strength	0.8	M4	0.50	1.00
Betweenness	0.8	M1	0.50	0.70
Betweenness	0.8	M2	0.50	0.68
Betweenness	0.8	M3	0.50	1.00
Betweenness	0.8	M4	0.50	1.00

Table 2. Convergence rates of models using different imputation approaches for various parameter combinations. M1: partial network - cross-sectional imputation; M2: partial network – longitudinal imputation; M3: full network – cross-sectional imputation; M4: full network – longitudinal imputation.

Network measure	**Model**	**Group capture probability**	**Convergence rate**
Strength	M1	0.25	0.98
Strength	M1	0.50	0.99
Strength	M1	0.75	1.00
Strength	M2	0.25	1.00
Strength	M2	0.50	1.00
Strength	M2	0.75	1.00
Strength	M3	0.25	1.00
Strength	M3	0.50	1.00
Strength	M3	0.75	1.00
Strength	M4	0.25	1.00
Strength	M4	0.50	1.00
Strength	M4	0.75	1.00
Betweenness	M1	0.25	0.65
Betweenness	M1	0.50	0.94
Betweenness	M1	0.75	0.98
Betweenness	M2	0.25	0.96
Betweenness	M2	0.50	1.00
Betweenness	M2	0.75	1.00
Betweenness	M3	0.25	0.96
Betweenness	M3	0.50	0.97
Betweenness	M3	0.75	0.99
Betweenness	M4	0.25	1.00
Betweenness	M4	0.50	1.00
Betweenness	M4	0.75	1.00

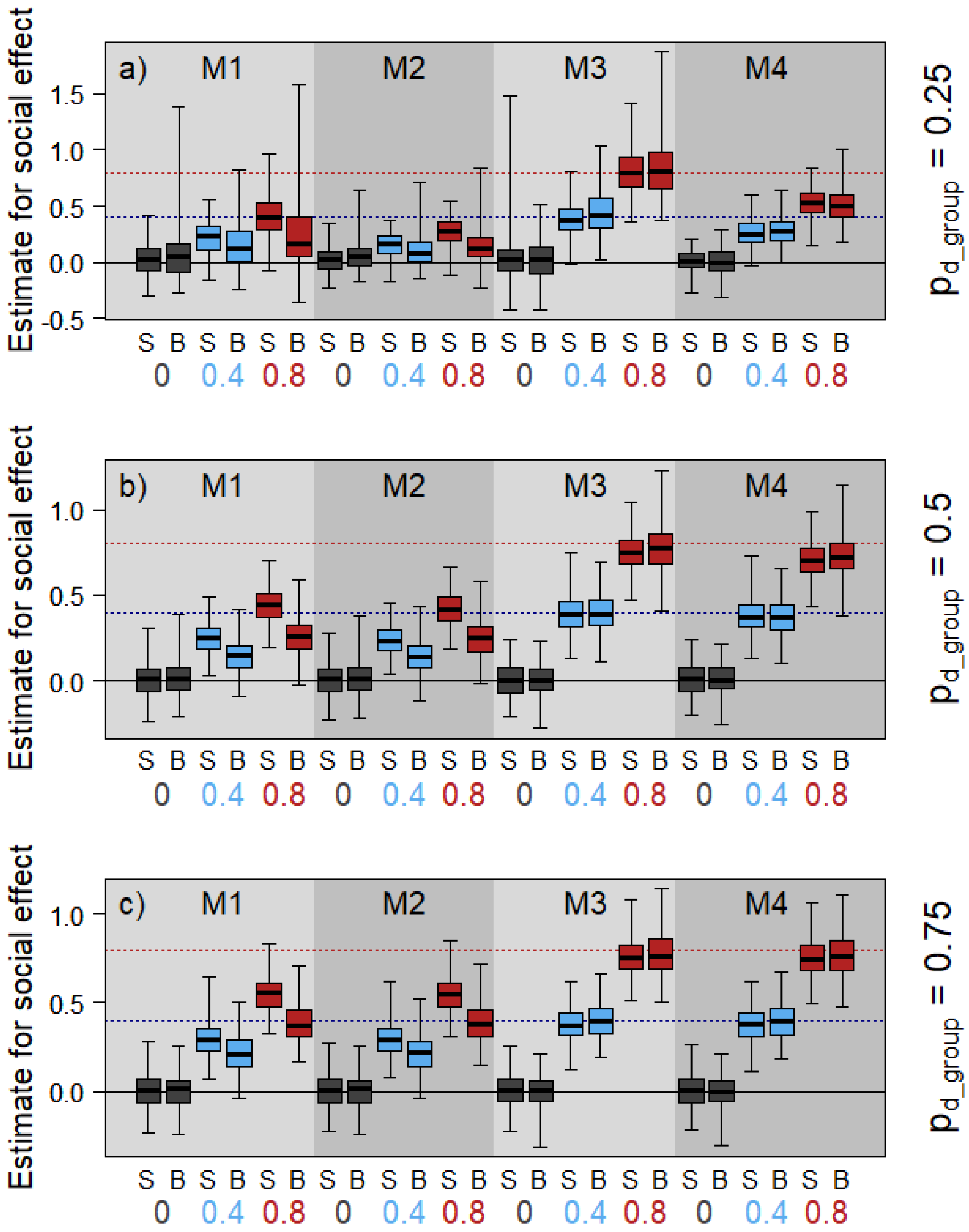


Figure 3. Distribution of posterior medians for the social effect of survival for different combinations of model (shaded polygons; M1: partial network - cross-sectional imputation; M2: partial network – longitudinal imputation; M3: full network – cross-sectional imputation; M4: full network – longitudinal imputation), network measure (S=Strength; B=Betweenness) and true effect size (box colour) when a) 25% of groups are sampled, b) 50% of groups are sampled

and c) 75% of groups are sampled. The solid central line represents the median, boxes the interquartile range and whiskers the full range of values.

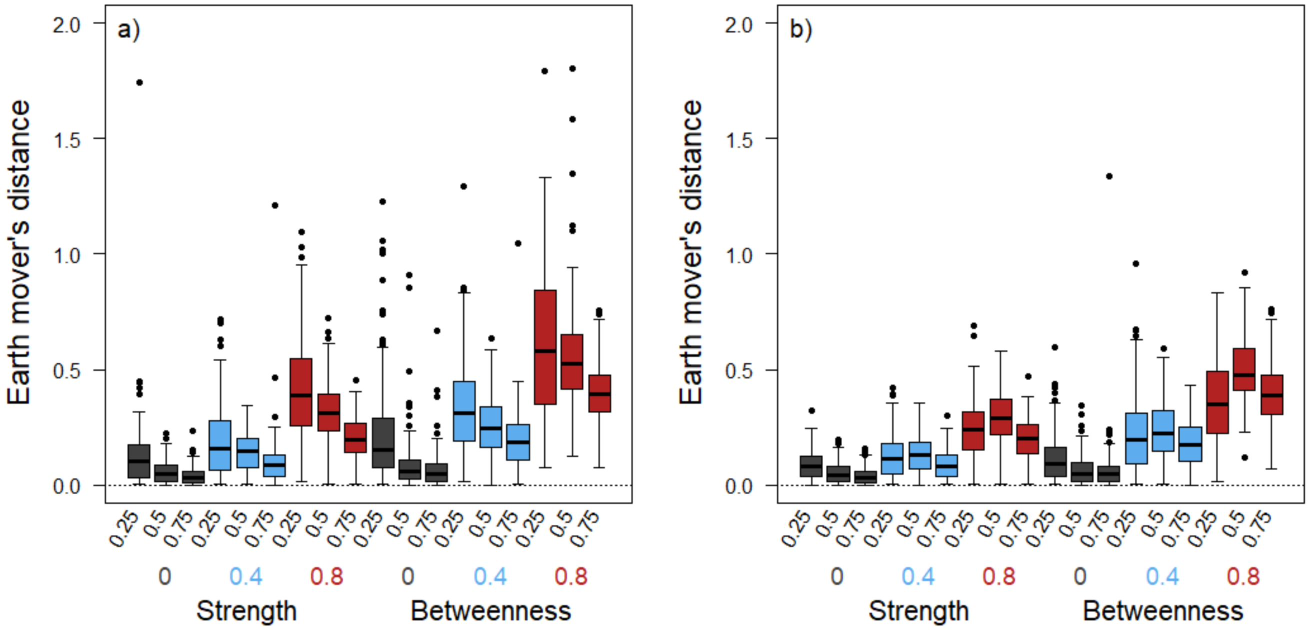


Figure 4. Earth mover's distances demonstrating the similarity of the posterior distributions for model estimates of the social effect on survival between the model using full and partial network measures for a) cross-sectional and b) longitudinal imputation. The solid central line represents the median, boxes the interquartile range and whiskers extend to 1.5x the interquartile range. We show the distributions for different combinations of group capture probability (0.25, 0.5 or 0.75), true simulated effect size (grey for 0, blue for 0.4 and red for 0.8) and the network measure influencing survival probability (strength vs betweenness). To aid visualisation we have excluded 19 outlying points with EMD>2 for panel a).

Case study 2: Effective sampling strategies to estimate social effects on survival

Specific Methods

Overview of Data generation

We simulated a population of 200 individuals with either a) no underlying group structure; or b) divided into 20 groups. Individual variation in the population was restricted to a single two-level categorical variable – sex. Underlying network structure depended on parameter choice (see below).

We simulated the behaviour and survival of individuals over 10 demographic timesteps, each of which contained 20 behavioural timesteps. As previously, grouping events had a mean size of two individuals. The survival probability of each individual depended on its sex (fixed effect of 0.5 on a logit scale) and position in the social network calculated from grouping events (see below) with a baseline survival probability of 0.8 in females. In this case study, there was recruitment into the population over time (i.e. the population stayed roughly constant over each simulation). There was a 10% chance that a surviving individual rewired its underlying social connections and if it did each connection had a 50% chance of changing.

The population was initially unmarked. Captures only occurred in the first behavioural timestep of each demographic timestep with 90% of groups sampled and a 0.9 probability of individuals in a sampled group being detected. Sampling design and effort for subsequent observations depended on parameter choice (see below).

Simulation structure

In total we generated 2880 simulated datasets, varying five parameters that influenced network structure (one parameter), network effects on survival (two parameters) and sampling effort/design (two parameters).

- Network structure: we varied underlying network structure so that either a) there was no group structure and moderate spatial structure driving the probability and weight of edges; or b) the population was divided into 20 groups with the probability of a within-group connection of 0.5 and within-group connection weights having a mean of 0.5 (versus a baseline of 0.2 and 0.25 respectively for between-group connections prior to adjusting for distance effects).
- Network effects on survival: a) we varied the network measure that influenced survival to be either strength or betweenness; and b) we varied the effect size to be 0 (no effect), 0.4 (moderate effect) or 0.8 (strong effect).
- Sampling: a) we varied sampling design so that the probability of observing a group within a sampled behavioural timestep covaried with the number of behavioural timesteps sampled in each demographic timestep resulting in (approximately) equivalent sampling effort being divided over the full demographic timestep. The probability of capturing a group was either 0.1, 0.2, 0.4 or 1, with the number of behavioural timesteps observed being 19, 10, 5 or 2; b) we varied the probability of an individual being observed in a sampled group to be either 0.5, 0.75 or 1.

For each combination of parameters (144) we ran 20 replicates.

Model-fitting

Unlike Case study 1 each Cormack-Jolly-Seber model was conditioned on first capture (as individuals were not assumed to have been captured previously).

Results and Discussion

Overall, survival models performed adequately in detecting social effects on survival (Table 3, Fig. 5, Tables S7-9, Fig. S7). When we simulated positive effects of network centrality on survival probabilities model estimates reflected this, although were substantial underestimates, especially with only moderate social effects on survival. These results support those from Case Study 1 indicating that it is possible to estimate social effects on survival, but that statistical power is limited with the presence of non-detected individuals and/or when many interaction events are unobserved. More encouragingly we show that for two very different social network structures there is little evidence for strong bias or elevated false positive rates when there is no social effect on survival.

Network variable

Our statistical models were better able to detect the effect of strength (local centrality measure) than betweenness (global centrality measure) on survival probabilities. While, the effect size was underestimated for both measures, this bias was much greater for betweenness centrality (Fig. 5), and results were more frequently statistically unclear (Table 3). The results here support those from Case Study 1 and the existing literature (Silk et al., 2015; Smith & Moody, 2013) in highlighting that global measures of network position are more susceptible to sampling effects than local measures.

Sampling design

There was no clear effect of how groups were sampled within each demographic timestep on estimates of social effects on survival (Fig. 5). Unsurprisingly, probability of observing individuals within groups did have some effect, with less downward-biased parameter estimates and more statistical power when sampling within groups was more complete (Fig. S7, Tables S7-8), as would be expected.

Lower observation success within sampled groups leading to reduced model performance is unsurprising as it leads to missing edges in the sampled network, reducing its correlation with the true (unobserved) network. This finding supports related work focussed on calculating network measures (e.g. (Franks et al., 2010)). (Franks et al., 2010) also tentatively recommended that more censuses (behavioural timesteps sampled in our case) were preferable than ensuring a high proportion of interaction events sampled in each census for calculating weighted measures of centrality. However, we found no clear evidence that this extended to our survival analysis, where there were only small differences in model performance and no clear overall trend. It should be noted, however, that the simulation architecture differed between the two papers.

Social structure

Social structure had a small effect on the ability to detect social effect on survival, with some differences in statistical power between the two structures investigated. While there were minimal differences in posterior medians (Fig. 5), results tended to be statistically clearer when there was no underlying group structure than when the population was divided into 20 groups (Table 3). Previous studies of sampling in social networks have rarely considered the types of modular social structures common for group-living animal populations (Silk, 2018). The slight negative impact of this group-structure on our ability to detect social effects on survival perhaps suggests that the correlation between network measures calculated in the sampled and full networks is weaker in these types of networks.

Estimates of other parameters

Estimates of other parameters were largely unaffected by social effects or sampling design. Strong social effects on survival were associated with slightly lower estimates of mean survival probability, but these differences were caused by differences in simulated survival probabilities rather than model performance (Figs. S8-10). While limited in scope these results provide evidence that including social effects on survival in demographic models is unlikely to impact other parameter estimates substantially (see also (Clements et al., 2022)).

Table 3. Proportion of simulation runs where 0 falls outside the 89% HDI for different parameter combinations with the probability of within-group detection fixed at 1.

Network measure	**True effect**	**Social structure**	**Sampling design**	**Detection rate**
Strength	0.4	Communities	0.1	0.45
Strength	0.4	No communities	0.1	0.50
Strength	0.4	Communities	0.2	0.40
Strength	0.4	No communities	0.2	0.70
Strength	0.4	Communities	0.4	0.55
Strength	0.4	No communities	0.4	0.70
Strength	0.4	Communities	1.0	0.50
Strength	0.4	No communities	1.0	0.60
Betweenness	0.4	Communities	0.1	0.20
Betweenness	0.4	No communities	0.1	0.10
Betweenness	0.4	Communities	0.2	0.00
Betweenness	0.4	No communities	0.2	0.10
Betweenness	0.4	Communities	0.4	0.00
Betweenness	0.4	No communities	0.4	0.15
Betweenness	0.4	Communities	1.0	0.10
Betweenness	0.4	No communities	1.0	0.15
Strength	0.8	Communities	0.1	0.95
Strength	0.8	No communities	0.1	1.00
Strength	0.8	Communities	0.2	1.00
Strength	0.8	No communities	0.2	1.00
Strength	0.8	Communities	0.4	0.95
Strength	0.8	No communities	0.4	1.00
Strength	0.8	Communities	1.0	1.00
Strength	0.8	No communities	1.0	1.00
Betweenness	0.8	Communities	0.1	0.15
Betweenness	0.8	No communities	0.1	0.35
Betweenness	0.8	Communities	0.2	0.45
Betweenness	0.8	No communities	0.2	0.45
Betweenness	0.8	Communities	0.4	0.25
Betweenness	0.8	No communities	0.4	0.40
Betweenness	0.8	Communities	1.0	0.40
Betweenness	0.8	No communities	1.0	0.50

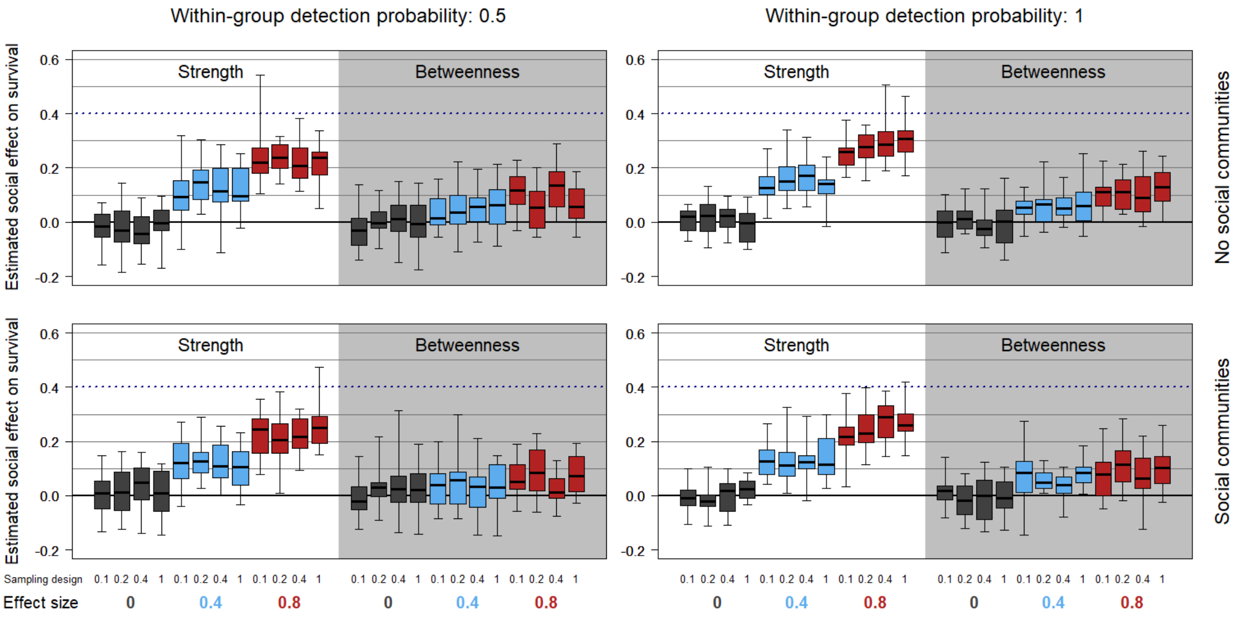


Figure 5. The impacts of sampling design (within-plot: sets of boxes of the same colour), within-group detection probability (columns) and social structure (rows) on Cormack-Jolly-Seber estimates of social effects on survival probability for a range of simulated effect sizes (colours of boxes). Boxplots show the distribution of posterior medians from multiple simulation runs with the solid line the median, boxes the interquartile range and whiskers the full range of values. We illustrate contexts in which a local measure of centrality (strength) and global measure of centrality (betweenness) are used as explanatory variables. The blue-dotted line indicates the accurate parameter estimate when the true effect size is 0.4 (the equivalent line for 0.8 is not illustrated).

Future steps

With the two case studies presented we can only scratch the surface of the potential of genNetDem as a methodological tool for animal social network analyses. Below we highlight some logical next steps for methodological studies on this topic, focussing on the integration of social networks and demography.

First, while we demonstrated the capacity for genNetDem to generate diverse social structures (Fig. 1), this was only a partial focus of our results. Animal social systems vary widely, and while sampling effects optimal sampling strategies are likely to vary with social structure (Clements et al., 2022; Silk, 2018; Sunga et al., 2021), this has remained understudied. Similarly, while we varied network dynamics in our simulations, the network of every individual was drawn from the same probability distribution. Incorporating greater trait-based or individual variation in network position would likely influence conclusions drawn about imputation approaches, for example.

Second, it is clear that simple approaches to imputing missing network measures are only partially successful; while they successfully generate qualitatively correct results, parameter estimates for social effects on survival are underestimated. While developing more sophisticated approaches to impute values for network measures may help, exploiting recent developments in network imputation (R. W. Krause et al., 2018, 2020; Young et al., 2020) are likely to have the greatest success. The adaptation of these novel approaches for behavioural ecology, and specifically within this capture-recapture modelling framework is a key challenge. (Gimenez et al., 2019) applied basic network imputation to study the social structure of Commerson's dolphin *Cephalorhynchus commersonii*. Similarly, (Clements et al., 2022) et al. included estimation of network structure within a Cormack-Jolly-Seber model to improve estimation of social effects on survival. However, the latter approach used a rather basic generative model for the latent network structure that could be improved on or adjusted for researchers working in different contexts. Consequently, extending these approaches to incorporate more sophisticated social network models as well as to open populations is a key priority.

Third, to keep our case studies accessible we examined social effects only in Cormack-Jolly-Seber models to estimate survival probability. (Clements et al., 2022) highlighted the potential value of incorporating social networks within integrated population models (IPMs), where different data sources could also be used to inform network structure itself. However, especially with improvements to imputation of latent network structures, there is also great potential to incorporate network effects within multi-state models more generally. Given the central role of social behaviour in mediating interactions between infectious disease dynamics and demographic processes (Silk et al., 2019; Silk & Fefferman, 2021), extending multistate models to incorporate social network structure in this way could provide important new insights into wildlife disease ecology, to provide just one example. genNetDem can provide an ideal sandbox to refine these models for application to wild systems.

Finally, we focus here on dyadic social networks, however many of the social interactions studied are non-dyadic and may include higher-order interactions (Battiston et al., 2021; Greening Jr et al., 2015). While there has been limited focus on higher-order interactions in animal societies (Musciotto et al., 2022), theory suggests they will impact infectious disease transmission and social contagions (Battiston et al., 2021; Iacopini et al., 2022; Noonan & Lambiotte, 2021) among other ecological and evolutionary processes. Therefore, expanding some of the developments here beyond dyadic networks to consider higher-order effects on survival and imputation of hyperedges (social connections between more than two individuals) will likely represent valuable developments. Because it generates GBIs that incorporate interactions/associations between more than two individuals genNetDem is an ideal starting point for methodological research testing higher-order methods in animal societies.

Conclusions

We introduce the R package genNetDem as a flexible tool for simulating combined social and demographic datasets. While we focus on the integration of social network and demographic models, the modular design of the package allows it to be an equally powerful tool for generating social network or capture-recapture datasets in their own right. It therefore provides a general tool for researchers interested in testing key methodological considerations in animal social network studies, especially as the field moves towards longitudinal analysis. It also helps researchers wishing to test the power of specific analyses or sampling designs in their own study systems.

Acknowledgements

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No. 101023948. We thank Sarah Clements, Qing Zhao, Mitch Weegman and Dave Hodgson for insightful discussions related to this work.

Data/Code availability

Data and code are available at https://github.com/matthewsilk/NETDEM and will be archived on acceptance.

References

Battiston, F., Amico, E., Barrat, A., Bianconi, G., Ferraz de Arruda, G., Franceschiello, B., Iacopini, I., Kéfi, S., Latora, V., & Moreno, Y. (2021). The physics of higher-order interactions in complex systems. *Nature Physics*, *17*(10), 1093–1098.

Blumstein, D.T., Williams, D.M., Lim, A.N., Kroeger, S., & Martin, J.G.A. (2018). Strong social relationships are associated with decreased longevity in a facultatively social mammal. *Proc. R. Soc. B*, *285*(1871), 20171934.

Butts, C.T. (2014). *Package 'sna.'*

Clements, S.J., Zhao, Q., Silk, M.J., Hodgson, D.J., & Weegman, M.D. (2022). Modelling associations between animal social structure and demography. *Animal Behaviour*, *188*, 51–63.

Croft, D.P., Madden, J.R., Franks, D.W., & James, R. (2011). Hypothesis testing in animal social networks. *Trends in Ecology & Evolution*, *26*(10), 502–507.

Csardi, G., & Nepusz, T. (2006). The igraph software package for complex network research. *InterJournal, Complex Systems*, *1695*(5), 1–9.

de Valpine, P., Paciorek, C., Turek, D., Michaud, N., Anderson-Bergman, C., Obermeyer, F., Wehrhahn Cortes, C., Rodrìguez, A., Temple Lang, D., & Paganin, S. (2022). *{NIMBLE}: {MCMC}, Particle Filtering, and Programmable Hierarchical Modeling*. https://doi.org/10.5281/zenodo.1211190

de Valpine, P., Turek, D., Paciorek, C., Anderson-Bergman, C., Temple Lang, D., & Bodik, R. (2017). Programming with models: writing statistical algorithms for general model structures with {NIMBLE}. *Journal of Computational and Graphical Statistics*, *26*(2), 403–413. https://doi.org/10.1080/10618600.2016.1172487

Ellis,S., Franks, D.W., Nattrass, S., Cant, M.A., Weiss, M.N., Giles, D., Balcomb, K.C., & Croft, D.P. (2017). Mortality risk and social network position in resident killer whales: sex differences and the importance of resource abundance. *Proc. R. Soc. B*, *284*(1865), 20171313.

Evans, J., Fisher, D.N., & Silk, M.J. (2020). The performance of permutations and exponential random graph models when analyzing animal networks. *Behavioral Ecology*, 31(5), 1266-1276.

Farine, D.R. (2013). Animal social network inference and permutations for ecologists in R using asnipe. *Methods in Ecology and Evolution*, *4*(12), 1187–1194.

Farine, D.R., & Carter, G.G. (2020). Permutation tests for hypothesis testing with animal social data: problems and potential solutions. *BioRxiv*.

Fisher, D.N., Ilany, A., Silk, M.J., & Tregenza, T. (2017). Analysing animal social network dynamics: the potential of stochastic actor-oriented models. *Journal of Animal Ecology*, *86*(2). https://doi.org/10.1111/1365-2656.12630

Formica, V.A., Wood, C.W., Larsen, W.B., Butterfield, R.E., Augat, M.E., Hougen, H.Y., & Brodie, E.D. (2012). Fitness consequences of social network position in a wild population of forked fungus beetles (Bolitotherus cornutus). *Journal of Evolutionary Biology*, *25*(1), 130–137.

Franks, D.W., Ruxton, G.D., & James, R. (2010). Sampling animal association networks with the gambit of the group. *Behavioral Ecology and Sociobiology*, *64*(3), 493–503.

Gimenez, O., Lebreton, J.-D., Gaillard, J.-M., Choquet, R., & Pradel, R. (2012). Estimating demographic parameters using hidden process dynamic models. *Theoretical Population Biology*, *82*(4), 307–316. https://doi.org/https://doi.org/10.1016/j.tpb.2012.02.001

Gimenez, O., Mansilla, L., Klaich, M.J., Coscarella, M.A., Pedraza, S.N., & Crespo, E.A.

(2019). Inferring animal social networks with imperfect detection. *Ecological Modelling*, *401*, 69–74.

Greening Jr, B.R., Pinter-Wollman, N., & Fefferman, N.H. (2015). Higher-Order Interactions: Understanding the knowledge capacity of social groups using simplicial sets. *Current Zoology*, *61*(1), 114–127. https://doi.org/10.1093/czoolo/61.1.114

Iacopini, I., Petri, G., Baronchelli, A., & Barrat, A. (2022). Group interactions modulate critical mass dynamics in social convention. *Communications Physics*, *5*(1), 1–10.

Krause, J., James, R., Franks, D.W., & Croft, D.P. (2014). *Animal social networks*. Oxford University Press.

Krause, R.W., Huisman, M., & Snijders, T.A.B. (2018). Multiple imputation for longitudinal network data. *Statistica Applicata-Italian Journal of Applied Statistics*, *1*, 33–57.

Krause, R.W., Huisman, M., Steglich, C., & Snijders, T. (2020). Missing data in cross-sectional networks–An extensive comparison of missing data treatment methods. *Social Networks*, *62*, 99–112.

Lebreton, J.-D., Burnham, K.P., Clobert, J., & Anderson, D.R. (1992). Modeling survival and testing biological hypotheses using marked animals: a unified approach with case studies. *Ecological Monographs*, *62*(1), 67–118.

Lee, C., & Wilkinson, D.J. (2019). A review of stochastic block models and extensions for graph clustering. *Applied Network Science*, *4*(1), 122. https://doi.org/10.1007/s41109-019-0232-2

McClintock, B.T., Langrock, R., Gimenez, O., Cam, E., Borchers, D.L., Glennie, R., & Patterson, T.A. (2020). Uncovering ecological state dynamics with hidden Markov models. *Ecology Letters*, *23*(12), 1878–1903.

Musciotto, F., Papageorgiou, D., Battiston, F., & Farine, D.R. (2022). Beyond the dyad:

uncovering higher-order structure within cohesive animal groups. *BioRxiv*.

Noonan, J., & Lambiotte, R. (2021). Dynamics of majority rule on hypergraphs. *Physical Review E*, *104*(2), 24316.

Oh, K.P., & Badyaev, A.V. (2010). Structure of social networks in a passerine bird: consequences for sexual selection and the evolution of mating strategies. *The American Naturalist*, *176*(3), E80–E89.

Opsahl, T. (2009). *Structure and Evolution of Weighted Networks*. University of London (Queen Mary College). http://toreopsahl.com/publications/thesis/

Pinter-Wollman, N., Hobson, E.A., Smith, J.E., Edelman, A.J., Shizuka, D., de Silva, S., Waters, J.S., Prager, S.D., Sasaki, T., & Wittemyer, G. (2013). The dynamics of animal social networks: analytical, conceptual, and theoretical advances. *Behavioral Ecology*, art047.

Pradel, R. (2005). Multievent: An Extension of Multistate Capture–Recapture Models to Uncertain States. *Biometrics*, *61*(2), 442–447. https://doi.org/https://doi.org/10.1111/j.1541-0420.2005.00318.x

R Core Team. (2021). *R: A Language and Environment for Statistical Computing*.

Shizuka, D., & Johnson, A.E. (2020). How demographic processes shape animal social networks. *Behavioral Ecology*, *31*(1), 1–11.

Silk, M.J. (2018). The next steps in the study of missing individuals in networks: a comment on Smith et al. (2017). *Social Networks*, *52*. https://doi.org/10.1016/j.socnet.2017.05.002

Silk, M.J., Croft, D.P., Delahay, R.J., Hodgson, D.J., Weber, N., Boots, M., & McDonald, R.A. (2017). The application of statistical network models in disease research. *Methods in Ecology and Evolution*, *8*(9). https://doi.org/10.1111/2041-210X.12770

Silk, M.J., & Fefferman, N.H. (2021). The role of social structure and dynamics in the maintenance of endemic disease. *Behavioral Ecology and Sociobiology*, *75*(8), 1–16.

Silk, M.J., Hodgson, D.J., Rozins, C., Croft, D.P., Delahay, R.J., Boots, M., & McDonald, R.A. (2019). Integrating social behaviour, demography and disease dynamics in network models: applications to disease management in declining wildlife populations. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *374*(1781), 20180211. https://doi.org/10.1098/rstb.2018.0211

Silk, M.J., Jackson, A.L., Croft, D.P., Colhoun, K., & Bearhop, S. (2015). The consequences of unidentifiable individuals for the analysis of an animal social network. *Animal Behaviour*, *104*, 1–11. https://doi.org/10.1016/j.anbehav.2015.03.005

Silk, M.J., McDonald, R.A., Delahay, R.J., Padfield, D., & Hodgson, D.J. (2021). CMRnet: An r package to derive networks of social interactions and movement from mark–recapture data. *Methods in Ecology and Evolution*, *12*(1), 70–75.

Smith, J.A., & Moody, J. (2013). Structural effects of network sampling coverage I: Nodes missing at random. *Social Networks*, *35*(4), 652–668.

Snijders, L., Blumstein, D.T., Stanley, C.R., & Franks, D.W. (2017). Animal social network theory can help wildlife conservation. *Trends in Ecology & Evolution*, *32*(8), 567–577.

Stanton, M.A., & Mann, J. (2012). Early social networks predict survival in wild bottlenose dolphins. *PloS One*, *7*(10), e47508.

Sunga, J., Webber, Q.M.R., & Broders, H.G. (2021). Influence of number of individuals and observations per individual on a model of community structure. *Plos One*, *16*(6), e0252471.

Touzalin, F., Petit, E.J., Cam, E., Stagier, C., Teeling, E.C., & Puechmaille, S.J. (2022). Mark loss can strongly bias demographic rates in multi-state models: a case study with simulated and empirical datasets. *BioRxiv*.

Young, J.-G., Cantwell, G.T., & Newman, M.E.J. (2020). Bayesian inference of network structure from unreliable data. *Journal of Complex Networks*, *8*(6), cnaa046.