Permutation tests for hypothesis testing with animal social data: problems and potential solutions

Damien R. Farine[1,2,3] and Gerald G. Carter[4,5]

[1]Department of Collective Behavior, Max Planck Institute of Animal Behavior, Konstanz, Germany.
[2]Centre for the Advanced Study of Animal Behaviour, University of Konstanz, Germany.
[3]Department of Biology, University of Konstanz, Germany.
[4]Department of Ecology, Evolution, and Organismal Biology, The Ohio State University, Columbus, USA
[5]Smithsonian Tropical Research Institute, Balboa, Ancon, Panama

Correspondance: dfarine@ab.mpg.de

ABSTRACT

1. Generating insights about a null hypothesis requires not only a good dataset, but also statistical tests that are reliable and actually address the null hypothesis of interest. Recent studies have found that permutation tests, which are widely used to test hypotheses when working with animal social network data, can suffer from high rates of type I error (false positives) and type II error (false negatives).
2. Here, we first outline why pre-network and node permutation tests have elevated type I and II error rates. We then propose a new procedure, the double permutation test, that addresses some of the limitations of existing approaches by combining pre-network and node permutations.
3. We conduct a range of simulations, allowing us to estimate error rates under different scenarios, including errors caused by confounding effects of social or non-social structure in the raw data.
4. We show that double permutation tests avoid elevated type I errors, while remaining sufficiently sensitive to avoid elevated type II errors. By contrast, the existing solutions we tested, including node permutations, pre-network permutations, and regression models with control variables, all exhibit elevated errors under at least one set of simulated conditions. Type I error rates from double permutation remain close to 5% in the same scenarios where type I error rates from pre-network permutation tests exceed 30%.
5. The double permutation test provides a potential solution to issues arising from elevated type I and type II error rates when testing hypotheses with social network data. We also discuss other approaches, including restricted node permutations, testing multiple null hypotheses, and splitting large datasets to generate replicated networks, that can strengthen our ability to make robust inferences. Finally, we highlight ways that uncertainty can be explicitly considered during the analysis using permutation-based or Bayesian methods.

INTRODUCTION

Permutation tests are among the most useful statistical tools for the modern biologist. They are commonly used in ecology (Gotelli & Graves, 1996), biogeography (Harvey, 1987), community ecology (Miller, Farine, & Trisos, 2017), and in studies of ecological networks (Dormann, Fründ, Blüthgen, & Gruber, 2009) and social networks (Croft, Madden, Franks, & James, 2011). Permutation tests randomize (or re-assign) observed data with respect to particular features to generate a distribution of statistic values that would be expected under a given null hypothesis. They are particularly useful when the standard assumptions of other statistical tests are violated, as is the case with social network data. Perhaps most importantly, permutation tests enable the researcher to create case-specific null models by permuting data in specific ways (e.g. constraining permutations within specific groups) while keeping other aspects of the dataset the same (e.g. where and when observations were made). For example, to understand how social network structure differs from what would be expected if animals made random social decisions, a researcher can permute the observation data to create many expected networks that could have occurred in the absence of social preferences (*pre-network permutations*). Alternatively, a researcher can ask whether the distribution of trait attributes within a network is not random by preserving the same observed network properties in all the randomised networks and only permuting the node attributes (*node permutations*). The difference in the design of these two permutation approaches has important consequences. They make different assumptions, they have different strengths and weaknesses, and, ultimately, they assess different null hypotheses. These consequences, together with the diverse range of drivers of network structure, can make even the most basic test of a hypothesis using animal social networks surprisingly difficult.

Recent studies (Weiss et al. 2020; Puga-Gonzalez et al. 2020) have highlighted some of the challenges that researchers face when testing hypotheses about the relationship between a predictor and a response variable, when one of these variables is generated from social network data. Consider these common questions: Does pairwise kinship influence association rates (edge weights)? Does an individual's sex predict how many associates it has (degree centrality)? To test the null hypothesis in these cases, we need to know what the data would look like in the absence of the effect of interest (in these cases, no effect of kinship or sex, respectively). A node permutation test tells us whether there is a statistical relationship between the effect of interest and the patterns of connectivity among the nodes in the observed network. The problem, however, is that the animal social networks that we observe are often the outcome of many processes and effects, besides the hypothesized effect of interest, and a node permutation test is not able to distinguish the hypothesized effect from the contributions of other, confounding, effects. A pre-network permutation test can better control for a range of confounding effects, however, it is actually testing a different null hypothesis—that the network-generating processes is random.

We can broadly assign confounding effects—effects that contribute to structuring a social network but are unrelated to the hypothesis of interest—to two categories. First, there are multiple non-social "nuisance effects", such as biases in sampling, non-random spatial constraints, and temporal differences in the presence or absence of individuals. These nuisance effects constrain our ability to observe each individual or dyad equally, meaning that network edges represent the outcome of not only social decisions, but also methodological and other non-social processes or constraints. Nuisance effects are common to most empirical studies in ecology, but the impacts of nuisance effects can be particularly pronounced in social network studies because they estimate pairwise relationships among individuals, as opposed to measuring the individuals themselves, and relationships are much more numerous than individuals. Second, there can be multiple social effects, besides the effect of interest, that operate simultaneously to shape the connections among individuals. In other words, nonrandom social structure often has multiple social causes or constraints. For example, individuals might have a limited number of possible close associates or always vary in their tendency to associate with different groupmates. Another example is triadic closure: if an individual A is closely associated to individual B that is also closely associated to individual C, then this will necessarily result in more encounters between A and C even in the absence

of any social preferences between A and C. The relative roles of such social and non-social confounding effects on the observed network structure can be difficult to identify and disentangle (Figure 1). Failing to do so can easily lead to spurious outcomes (Farine & Aplin 2019).

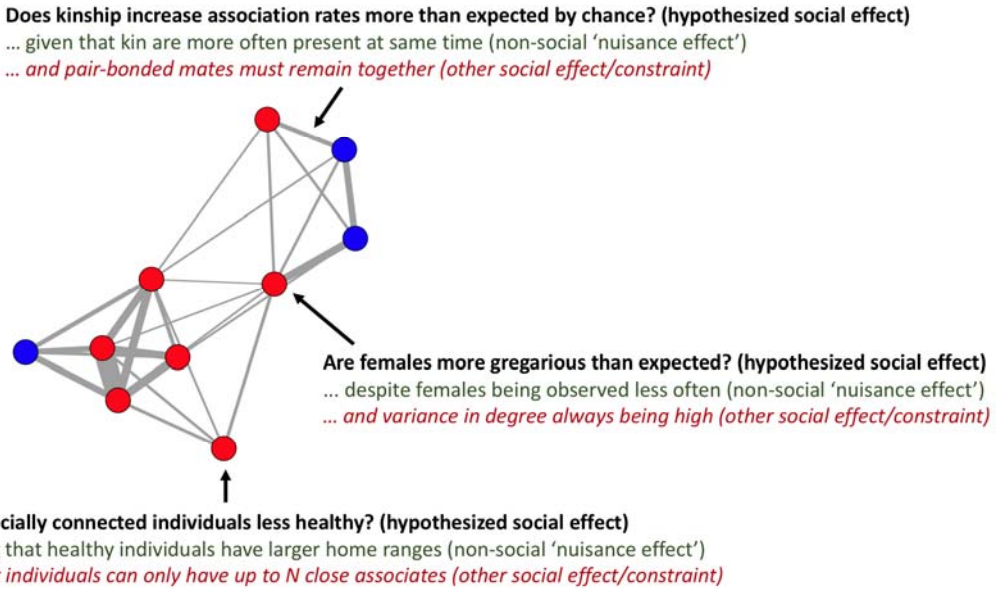


Figure 1. Observed social networks are usually the product of many different social and non-social effects, which can impact the observed difference or relationship that is expected by "chance". Researchers typically aim to test for a relationship between a measure taken from the social network data and some independently-measured data. Such tests can take the form of a predictor (e.g. sex, kinship, health) on a social response (e.g. degree centrality, edge weight, eigenvector centrality), of a social predictor (e.g. degree centrality) on a response (e.g. infection status), or via an estimation of the correlation between network and non-network-based measures. However, spurious relationships and correlations are common in social network data because of multiple non-social 'nuisance effects' (examples above) and other social effects or constraints (examples in italics above). We give examples of each effect above, but many other such effects are possible. Pre-network permutation tests can control for some non-social nuisance effects, but other social effects or constraints that shape the network structure are not maintained. Node permutation tests can control for the contribution of other social effects or constraints on the social network structure, but precisely controlling for non-social nuisance effects is more challenging.

Two common permutation approaches (pre-network permutations and node permutations) create null expectations that are better at addressing one category of confounding effects but not the other. Pre-network permutation tests swap observations to create a set of possible networks that would be expected from animals showing no social preferences (e.g. about which social groups to join), and these can effectively control for nuisance effects by constraining these swaps within blocks of time and space. For example, one might swap observations within sampled locations to control for spatial effects or within sampled time periods to control for non-overlapping presence and temporal autocorrelation in behaviour (Farine, 2017; Spiegel, Leu, Sih, & Bull, 2016; Sundaresan, Fischhoff, & Dushoff, 2009; Whitehead, 2008; Whitehead, Bejder, & Ottensmeyer, 2005). Unless explicitly designed to do so, pre-network permutation tests do not distinguish among alternative social processes, thereby creating a different problem. Pre-network permutation tests can yield unacceptably high rates (>30%) of false positives when drawing inferences about the effect of a predictor X on a response Y (i.e. linear models or difference between means, where one of X or Y is a network-based measure; Weiss et al. 2020; Puga-Gonzalez et al. 2020). These high type I error rates occur because pre-network permutation tests do not actually address the null hypothesis that X is distributed randomly with respect to Y (or that the effect of X on Y is zero); instead they test if the distribution of X with respect to Y is different than expected had individuals made random social choices given the possible options that were available to them. For example, spatially-restricted pre-

network permutations simulate a scenario where individuals' spatial decisions are not socially influenced, and all pre-network permutations assume that individuals make decisions independently of each other. In other words, because pre-network permutations aim to remove all bias from social decisions in the randomised networks, these networks also deviate from the realistic non-random social structures that are expected for a given species or population.

Node permutation tests face a different problem. Although they can test for a non-random relationship between variables X and Y in a way that controls for social structure, they also assume that the observed network corresponds to the real social structure (i.e. the structure based on social preferences in the absence of any non-social "nuisance effects"). This assumption makes sense if the social networks were completely accurate reflections of social preferences, but observed animal social networks (as with most biological data) are almost always shaped by at least some observational, spatial, and temporal biases. For example, some individual animals might only use a subset of all possible locations where observations were made, individuals might vary in the amount of their home range that overlaps with the study area, some individuals might leave or join the study population at different times, and some might not be individually marked or identifiable for the entire duration of the study. Even if some processes are relevant to the hypothesis of interest (for example individuals' decisions about where to settle in space) they can still contribute to some inaccuracies in the resulting network. For example, individuals at the edge of a study area (compared to individuals at the centre of the area) might have many more associations with individuals that were never observed (we discuss further examples below). These nuisance effects can vary in magnitude and importance across study designs, but they are arguably inevitable. Even automated methods such as proximity sensors (Ryder *et al.* 2012; Ripperger *et al.* 2020) or barcodes (Crall *et al.* 2015; Alarcón-Nieto *et al.* 2018) that aim to provide equal sampling across individuals would not be free of sampling biases, if animal-borne proximity sensors vary in their sensitivity (for example due to tiny differences in how they were soldered) or if some barcodes are more difficult to identify by computer vision. Thus, methodological factors are rarely completely eliminated, even under highly controlled conditions. Such sampling biases and other nuisance effects can lead to elevated rates of false positives and false negatives when using node permutations (Croft *et al.* 2011; Farine & Whitehead 2015; Farine 2017; Puga-Gonzalez, Sueur & Sosa 2020), and many can be quite difficult to correct using correction terms in a statistical model.

A major challenge, as it stands, is developing permutation methods that can robustly account for both social and non-social nuisance effects. One approach to dealing with nuisance effects is to control for them by including them as covariates or random effects in a statistical model. Incorporating a specific non-social nuisance effect, such as the number of observations of each individual (which affects network metrics like degree) explicitly into the model can correct the coefficient values. Doing so helps ensure that the zero value of a test statistic (like a t-value or linear slope) accurately represents the null hypothesis of interest, potentially alleviating the need for pre-network permutation tests (Franks et al. 2020). While there are major advantages to this approach, the process of capturing all nuisance effects in the model becomes increasingly challenging as the number of interacting effects increases. Observed network data often have multiple simultaneous nuisance effects, and attempting to measuring all of them individually, let alone in combination, can be more difficult than with a permutation-based approach, which maintain nuisance effects (and their variation) constant across the permuted networks. For example, it is challenging to assign individuals to a singular spatial location, as required when fitting individuals' location as a random effect, if home ranges are continuously distributed and overlapping in space. It is also difficult to control for cases where two individuals have both been repeatedly observed at the same location(s) but were never present there at the same time. A strength of pre-network permutation tests is that they can inherently control for multiple potential nuisance effects because the permuted data can be kept identical to the observed data with regards to the number of observations per individual, the size of groups, individual variation in space use (and therefore spatial overlaps with all others), temporal auto-correlation in behaviour, temporal overlap among all pairs of individuals, the distribution of demographic classes across space and time, the variation in the density of individuals

192 across space and time, differences in sampling effort across space and time, and the observability of
193 individuals.
194 Our goal here is to propose an initial solution to the current problems with the use of
195 permutation tests outlined above. Our solution uses both pre-network and node permutations for
196 what they each do best: pre-network permutations to control for nuisance effects, and node
197 permutations to statistically test for the effects of X on Y while holding the network structure
198 constant across expected networks. Our approach proceeds as follows: (1) we calculate the network-
199 based observations of interest (e.g. degree for each node or edge weight for each dyad), (2) we use
200 pre-network permutations to generate an expected null distribution of alternative network-based
201 expected values for each unit (node or edge), (3) to control for nuisance effects, we subtract the
202 median of the expected values for each unit from its corresponding observed value to create
203 residual-like estimates, (4) we fit these "residuals" into the statistical model, and (5) we use node
204 permutations to calculate the P value for the observed effect of X on Y, where the network-based
205 variable and corresponding test statistic are now corrected for nuisance effects.
206 This double permutation procedure tests the null hypothesis that deviations from random
207 social structure (within some set of constraints) are not explained by the predictor variables. The
208 procedure can be easily applied to any model for calculating test statistics, such as Mantel tests
209 (Mantel, 1967), network regression models like MRAQP (Dekker, Krackhardt, & Snijders, 2007), and
210 metrics such as the assortativity coefficient (Farine, 2014; Newman, 2002). As we will show, it also
211 performs equally well with group-based association data and with data collected using focal
212 observations. We acknowledge that this is only one potential solution, and we therefore also
213 highlight alternative methods that are also worth evaluating further, including several that are not
214 based on permutation tests.
215
216
217 **ILLUSTRATING THE DRIVERS OF TYPE I AND TYPE II ERRORS**
218
219 Before we discuss our solution in detail, let us clarify the main problem by considering a
220 simple, verbal, but concrete example of why errors can arise when using permutation tests. Imagine a
221 study population where animals cluster for warmth each night in variable groups of 2-10 individuals.
222 The dataset contains a list of observed clusters, their location, time, and the individuals in those
223 clusters that could be correctly identified. From these data, researchers generate a network
224 describing the propensity for each dyad to be observed in the same cluster with the aim of finding
225 out if kin are more likely to cluster. Specifically, they ask: Does the dyadic kinship predict the observed
226 propensity to be observed clustering together (edge weight)?
227 First, the researchers consider a node permutation approach (e.g. using a Mantel or MRQAP
228 test). However, if siblings are born at the same time, limited to similar home ranges, and then
229 disperse at around the same time—then they could be more associated with each other than with
230 non-kin, even without kin discrimination (e.g. Leedale *et al.* 2018). Under such a scenario, a
231 significant result from a node permutation would correctly support that the network is kin-
232 structured, but represent spurious support for the specific hypothesis that kinship is a driver of social
233 associations. Alternatively, a non-significant result may be caused by sampling bias. For instance,
234 associations among kin could be under-estimated if younger animals are both more likely to associate
235 with kin and less likely to be individually marked and recorded. In summary, hypotheses about the
236 process generating an effect of kinship on association could be challenging to accurately assess using
237 node permutations, in the presence of nuisance effects (but see Alternative Approaches section for
238 more discussion on how node permutations can be restricted to potentially help alleviate some of
239 these effects).
240 The researchers might therefore turn to using a pre-network permutation test. They create
241 expected outcomes (i.e. measure the relationship between kinship and edge weights) after
242 repeatedly swapping individual observations within sampled locations and time periods. Doing so
243 allows them to eliminate the nuisance effects described above. However, now imagine that there is
244 another unknown social effect: all individuals spend about 90% of their time with only 1-3 closely

bonded associates. This constraint on social structure means that every node (individual) always has a large variance in edge weights across all its possible associations, because many association rates are zeros while a few others are closer to one, and this variance is much greater than expected from random associations. When the researchers create randomised networks using pre-network permutations, the observations of individuals (i.e. their social actions) are swapped independently of each other, meaning that individuals in the resulting random networks almost never (or possibly never) spend 90% of their time with a few individuals, while many zeros are replaced with non-zeros as individuals that never co-occurred are swapped into groups together. While this removal of social preferences is one of the aims of pre-network permutation tests, it can confound non-random social structure generated by the trait of interest with that of other processes not related to the hypothesis being evaluated. In the context of the kinship scenario above, even if close social bonds are not kin-biased (i.e. the social bonds are randomly distributed with regards to kinship), the extremely non-random social structure of the observed network could easily lead to a false positive with regards to the effect of kinship. If even a few strong bonds exist among kin by chance in the observed data, it is possible that these strong kin bonds will never appear in the expected data. This scenario occurs because the observed network typically has much higher variance in the measure of interest (edge weight or node metric) than the corresponding random networks from pre-network permutations (Aplin *et al.* 2015; Firth *et al.* 2018; Weiss *et al.* 2020). Thus, once again, the actual effect of kinship on association is challenging to accurately evaluate in the presence of a confounding effect (in this case social preferences unrelated to kinship).

The false positive in the last example is related to a different potential problem with pre-network permutation tests—they can provide overly-confident estimates of minor deviation from random (Figure 2). In part, this problem occurs because constructing social networks requires large numbers of observations (Langen 1996; Farine & Strandburg-Peshkin 2015; Davis, Crofoot & Farine 2018) with many repeated observations of the same individuals, and a sufficiently large number of observations can invariably produce P values well below 0.05 even when the effect size is not biologically important (Figure 2).

Pre-network permutations are also likely to suffer from more incorrect inference in analyses based on networks containing only a few nodes. For instance, consider a network containing only three nodes (male, male and female), from which only three dyadic associations are recorded, with all three being between the two males. In such a scenario, a pre-network permutation testing whether females are less connected (lower degree) than expected by chance would be significant at $P<0.05$, as the observed data is the only one of 27 combinations of the three dyadic associations that would result in the female having a degree of zero (the P value here would approach 0.037). Clearly, any inference drawn from three observations of three individuals would be immediately apparent as unreliable, but the same problem can be harder to notice in more complex analyses.

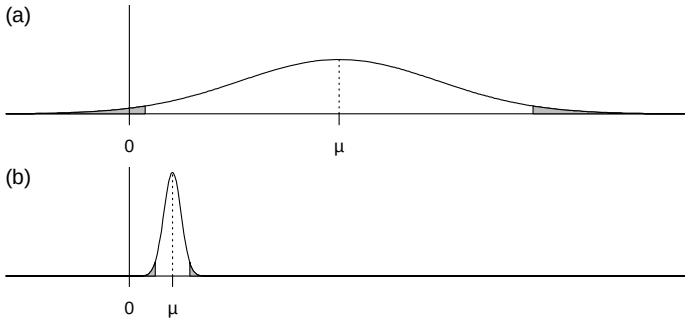


Figure 2. The problematic relationship between effect size and significance. (a) A large effect in a test with a relatively low power dataset, producing a *P* value of 0.016. (b) A weak effect in a test with a high-power dataset, producing a *P* value of 2.87×10^{-7}. The former is more biologically significant, whereas the latter is more statistically significant. When drawn from large numbers of observations, pre-network permutation tests can detect marginal differences that have little biological

relevance (i.e. b), thereby producing elevated type I errors (sensu Nakagawa & Cuthill 2007; Lantz 2013; Szucs & Ioannidis 2017).

The previous example illustrates problems caused by nuisance effects involves variables that are traits of dyads (edges), and the same principles apply when testing hypotheses relating the traits of individuals (nodes) to their connectivity in the network. For instance, a false impression that male birds are more gregarious (i.e. sex predicts degree centrality) could be caused by sex-based differences in observability, attraction to spatially clumped resources, variation in home-range size, differential survival in different habitats, or differences in the timing of their presence or absence in the population. Again, these issues are common to many other types of ecological studies, but their influence can be exacerbated in social network studies. Some of these effects are sampling effects; others will only cause a problem if there's a mismatch between the question of interest to the researcher and the question being addressed by the analysis (e.g. 'do males prefer larger groups?' versus 'do males occur in larger groups?', see Figure 3). Further, while we refer to many nuisance effects as 'non-social', social behaviours can also contribute to these effects, and studies could benefit from making more explicit considerations of the social decisions that contribute to them. For example, individuals' spatial ranges could be determined by social habitat selection (e.g. density-dependence in their decisions to settle in a location or move elsewhere). We provide some advice on how permutation tests can help uncover such effects in the future directions section.

Social behavior itself can affect observability. For example, if females tend to be found at the periphery of groups, then an observer standing close to the center of a group might be more likely to miss observations of females and their associates, whereas the observer can always detect the associates of the (predominately male) individuals at the center of a group, thereby introducing a sex bias in the number of observed associates. An observer standing outside the group could have the opposite bias. This simple example highlights how biases can arise even when observations are made in groups where every individual is individually-identifiable, and why the focus on detecting relationships (i.e. estimating edges) makes network studies more prone to nuisance effects than many other types of studies.

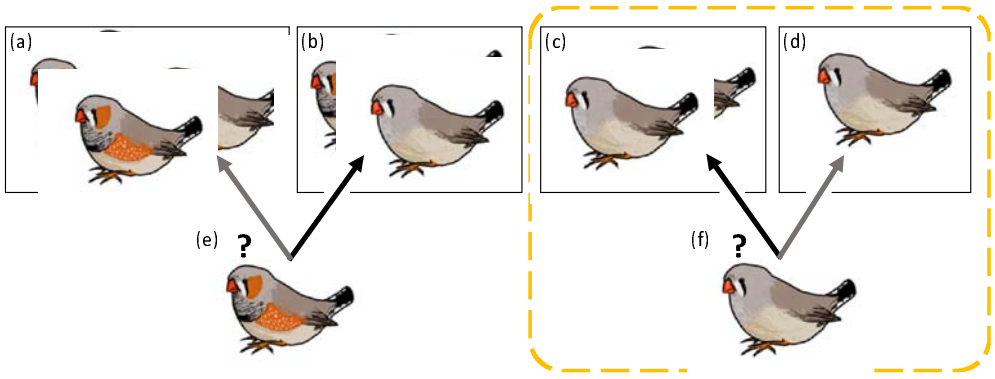


Figure 3. Example of the challenges associated with testing hypotheses on social data. Are males more often in larger groups? In the scenario shown above, males (colourful individuals) are observed in larger groups (e.g. groups a, b, c) more often than are females (less colourful individuals). However, now consider the question: Are males more gregarious than females? The same observation is not informative about the causal processes with respect to this question. Imagine that groups a and b represent birds outside a territory (shown by the dashed line) defended by a single dominant male (in c). Females are allowed to enter freely and form groups within the territory (c and d) but other males are excluded. Even if each male (e) chooses smaller groups (b vs a), and females (e.g. f) choose larger groups (c vs d) in their given environments, we could still observe males in larger groups than females due to the constraints imposed by the territory. Thus, even if males are less gregarious, they can end up having more social connections. A node permutation test (where the sex labels

are randomized in the overall network) that produced a significant result would incorrectly support for the hypothesis that males are more gregarious. By permuting the individual observations within locations (a pre-network permutation), we can simulate the random social decisions that individuals would make given the options that were available to them. Under such a null model, a subordinate male that was never seen in the territory (groups c and d) would never be observed there in the permuted data, and would therefore always be found in larger groups. Such a test would correctly avoid rejecting the null hypothesis that males are more gregarious (because in the randomised data they would always chose larger groups). However, the same pre-network permutation test would incorrectly fail to reject the null hypothesis that males and females do not differ in their numbers of connections, if that were the question of interest.

An advantage of pre-network permutations is that they allow precise control over many possible nuisance effects, without needing to measure or even identify them. For example, individuals at the edge of a study area, where a lower proportion of individuals are individually-identifiable, would always occur in groups containing fewer individuals that can be individually-identified; controlling for space would automatically control for differences in group size that arise due to spatial variation in identification rates. Such an effect could be challenging to control for by explicitly including it in a statistical model. However, pre-network permutation tests can allow many other important aspects of the data to change (such as degree distributions or variances) in the expected data, and thus pre-network permutations alone cannot be used to assess the effect of a predictor while controlling for social structure (Weiss et al. 2020, Puga-Gonzalez et al. 2020).

THE DOUBLE PERMUTATION METHOD

We propose an approach that uses pre-network permutations to control for nuisance effects, and then uses node permutations to test for the statistical significance of the effect of interest. Our double permutation testing method (Figure 4) first uses pre-network permutations to calculate the deviation of each of the units of interest (a node-level or edge-level metric) from its random expectation given the structure of the observation data. That is, by comparing a unit's observed measure to its expected random value (e.g. the median values of the same unit's measure across the permuted networks), we can calculate the equivalent of a residual value. These residual values can then be fit into a model of interest—such as an MRQAP, regression, or other model—to generate a corrected test statistic, and node permutations used to calculate the significance of this statistic. Such an approach is conceptually similar to generalised affiliation indices (Whitehead & James 2015), but it uses pre-network permutation tests, rather than regression models, to estimate the deviance from random, and it applies them directly to the metric of interest (e.g. a node's degree) rather than using a two-step process of calculating corrected affiliation indices before generating a given network.

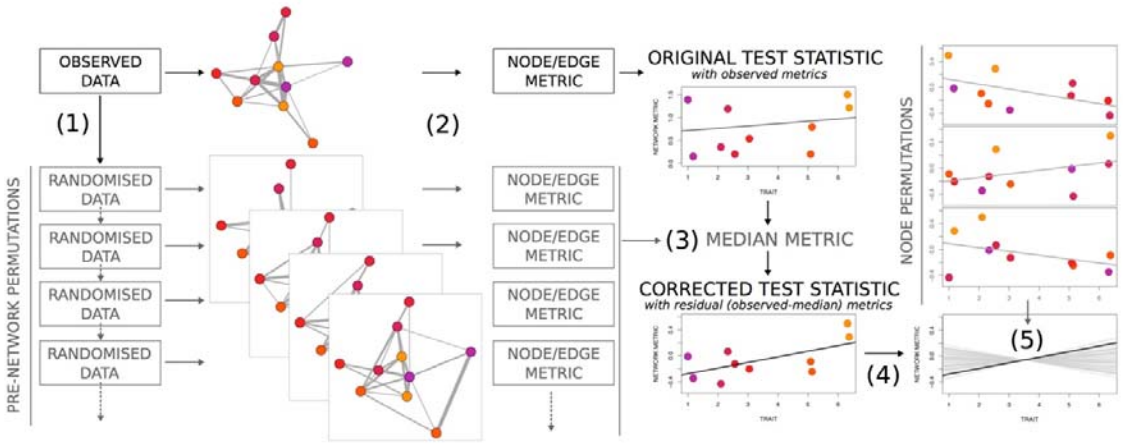


Figure 4. Overview of the double permutation method. We propose a solution to the problem of elevated type I and type II errors when using permutation tests in animal social network analysis. Our approach has four steps. (1) A pre-network permutation is used to (2) generate a distribution of expected metric values for a unit of interest (e.g. a node's degree or an edge's weight). For each unit (i.e. each node or each edge), (3) the unit's expected metric value (e.g. the median) is subtracted from its observed metric value, which yields a corrected metric value (the equivalent of residual values after controlling for non-social nuisance effects). (4) The test of interest (e.g. a regression, difference in means, or correlation), and its corresponding test statistic (e.g. the coefficient of the slope or the correlation coefficient) is calculated to generate a corrected test statistic. (5) A node permutation test, in which the trait values are shuffled relative to the residuals of the metric values, is then used to compare the corrected test statistic with those expected given the structure of the network, to generate a P value.

TESTING THE ROBUSTNESS OF THE DOUBLE PERMUTATION APPROACH

We demonstrate the suitability of our approach using three sets of simulations. In the first, we show that double permutation tests provide robust outputs when used with regression models on group-based data, both in the absence of any real relationship (to test for elevated type I error rates) and when there is a strongly confounding nuisance factor (e.g. a spatial effect, to test for elevated type II errors rates). In the second model, we use the same simulation framework as Puga-Gonzalez, Sueur and Sosa (2020) to demonstrate that double permutation tests are robust when using focal-observation data and when used to compare means. Finally, we develop a third model to show that double permutation tests are robust to testing edge-based hypotheses (e.g. the role of kinship in shaping the strength of connections among individuals) in the presence of other social effects (e.g. the presence of non-kin social bonds).

Simulation 1: node-based regression

The first simulation starts by drawing N individual trait values T_i from a normal distribution with a mean of 0 and a standard deviation of 2. We assign each individual to have on average K observations by drawing K_i from a Poisson distribution with $\lambda = K$ and balancing these values to ensure that $\sum K_i = N \times K$. We then create G groups, where $G = 0.5 \times N \times K$, and randomly assign each of these groups to have a group size value X ranging from 1 to 10. To allocate individuals into groups, we order the individuals from the smallest trait value to the largest to create scenarios where the trait value should impact the social behaviour of individuals (trait has social impact, $T_S = TRUE$), or order these at random to create scenarios where the trait value has no relation to the social

behaviour of individuals ($T_S = FALSE$). We assign each individual into groups by selecting the K_i groups that have empty spaces, and with a higher probability of selecting smaller groups. In doing so, individuals earlier in the order are disproportionately more likely to be assigned to smaller groups, filling them up, and leaving only larger groups for later individuals to fill, thereby creating a relationship between individuals' trait value T_i and their weighted degree D_i when T_S is true.

From these observation data, we follow the design of our double permutation method (see Figure 4) to first calculate how each node's degree deviates from what is expected from random behaviour, and, second, to calculate the relationship between the residual weighted degree values D_i' and the trait values T_i. We then implement two conceptual variations, L_S, to go with the two T_S scenarios above. The first variant (no location effect, $L_S = FALSE$) is a scenario in which the group size values X corresponds to the outcome of social decisions. In the second variant (location effect, $L_S = TRUE$), we assume that rather than X representing a group size preference, X instead corresponds to spatial preferences, such that individuals prefer patches closest to the centre of their home ranges in a one-dimensional linear environment ranging in values from 1 to 10. Patches at one end of this environment, i.e. those patches with a larger X, contain more resources and therefore can hold more individuals. These variants enable us to use the same code to produce a relationship between T_i and network degree D_i where in one scenario where the decisions are social ($T_S = TRUE$ and $L_S = FALSE$) and in another scenario where the relationship between T_i and D_i arises from decisions that are not social ($T_S = TRUE$ and $L_S = TRUE$). To control for the location X_j that each group was observed in when $L_S = TRUE$, we used within-location swaps in the pre-network permutation tests.

We simulate 100 replications for varying combinations of network sizes, with number of individuals ranging from $N = 5$ to $N = 120$ and for mean numbers of observations per individual ranging from $K = 5$ to $K = 40$. The different combinations of scenarios (T_S and L_S) allow us to evaluate the performance of different approaches in cases where there are no real effects ($T_S = FALSE$ and $L_S = FALSE$), real social effects but no nuisance effects ($T_S = TRUE$ and $L_S = FALSE$), and where the driving factor behind the effect is a strong non-social nuisance effect ($T_S = TRUE$ and $L_S = TRUE$). The latter represents an example of a study with a confound (differences in spatial preferences among individuals) that is challenging to control for because individuals are observed in most locations, meaning that their spatial preferences cannot be reduced down to a single value, which is required when model fitting (e.g. adding location as a random effect) or when restricting node permutations by location (because each node can only have one location attribute used for swapping).

For each run of the simulation, we calculate P values for the effect of the trait value on degree using (1) node permutation tests with the coefficient value as the test statistic, (2) node permutation tests with the coefficient value as the test statistic while controlling for number of observations (Franks *et al.* 2020), (3) pre-network permutation tests with the coefficient value as the test statistic, (4) pre-network permutation tests with the t statistic as the test statistic, and (5) double permutation tests with the coefficient as the test statistic. We extract β coefficients (slopes) and t statistics by fitting the model weighted degree (D) ~ trait (T) using the lm function in R. We create the networks and conduct the pre-network permutation tests using the R package *asnipe* (Farine 2013).

Simulation 2: difference in group means

We implement the second simulation using exactly the same code as Puga-Gonzalez, Sueur and Sosa (2020). In brief, these simulations start by assigning individuals to groups, with each group having a focal individual. Simulations can be run with and without a difference in gregariousness

among females and males, where females are more gregarious by being disproportionately allocated to larger groups when the effect is present. The simulations can also introduce an observation bias, whereby females are often not observed even when present, whereas males are always observed when present. Such biases are common in field studies—for example in a study on vulturine guineafowl (*Acryllium vulturinum*) (Papageorgiou *et al.* 2019), juveniles are marked with a soft wing tag on the right wing and therefore only identifiable if their right side is observable, whereas adults are marked with leg bands that can be identified from any direction. The code runs 500 simulations for each scenario (sex effect or not, observation bias present or not), with parameter values that are randomly drawn from uniform distributions as follows: population size ranging from 10 to 100, observation bias ranging from 0.5 to 1 (where 1 is always observed), the female sex ratio ranging from 0.2 to 0.8, and the number of focal follows ranging from 100 to 2000.

The simulation procedure above follows closely from the design in Farine (2017), and was designed to provide the ability to record both the pre-bias and post-bias effects, thereby allowing an estimation of false positives (when no effect should be present but one is detected), false negatives (when an effect is present, but masked by the observation bias, and therefore not detected), and whether the model can accurately estimate the original effect size (before the observation bias is applied). For each simulation, we calculate P values of the effect of sex on degree using (1) node permutation tests with the coefficient value as the test statistic, (2) pre-network permutation tests with the coefficient value as the test statistic, (3) pre-network permutation tests with the t statistic as the test statistic, and (4) double permutation tests with the coefficient as the test statistic. We extract β coefficients (the difference in the mean degree between males and females) and t statistics by fitting the model weighted degree ~ sex in the lm function in R.

Model 3: edge-based regression

We demonstrate that the double permutation method is robust to the presence of other social effects. To evaluate the impact of other social effects on error rates, we generate simulated networks in which individuals have three types of social associates: (i) weak associates, (ii) preferred associates, and (iii) strongly-bonded associates. We start (1) by creating a 'real' network comprised of N individuals with a network density D drawn from a uniform distribution (ranging from 0.05 to 0.6), and selecting $Z \times D$ edges (where Z is the maximum possible number of undirected edges) with probabilities 0.6, 0.3, and 0.1 for edge types 1 to 3, respectively (and all other edges set to 0). As with our first simulation, we then (2) make an average of K observations per individual. However, in the current simulation, we create $1.2 \times \max(K_i)$ sampling periods (Whitehead 2008), and randomly allocate individuals to being observed in K_i of these sampling periods. (3) For each sampling period, we then select all pairs of individuals with an edge present in the real network and where both are present in that sampling period, and draw a 0 or 1 to signify whether they were observed together or not. Here we set the binomial probability of drawing a 1 set to 0.1 for edges of weak associates, 0.6 for edges of preferred associates, and 0.9 for edges of strongly-bonded associates, based on the real network. These values therefore represent weak, strong, and very strong likelihoods of individuals being co-observed when they are both present. We select N and K values using the same parameter sets as in model 1.

Next, we (4) create a kinship network, setting the kinship level of each individual based on their edge type in the real network. Specifically, we draw relative kinship values from a beta distribution with $\alpha = 1$ and $\beta = 2$ (i.e. left-skewed) for missing edges and edges of weak associates, $\alpha = 2$ and $\beta = 2$ (i.e. unskewed) for edges of preferred associates, and $\alpha = 3$ and $\beta = 2$ (i.e. right-skewed) for edges of strongly-bonded associates. Because beta distributions range from 0 to 1, these

493 distributions assume that 1 corresponds to the closest relatives in the population. We use these
494 parameters to create kinship distributions with differences in mean kinship (0.33, 0.5, 0.6,
495 respectively) according to social relationship type (these relative kinship values can be divided by two
496 to create a maximum kinship value of 0.5 with no effect on the outputs of the model).
497 Our simulations comprise two scenarios. In the first scenario, edge weights are purely social
498 and unrelated to kinship, which we achieve by randomizing the kinship matrix relative to the
499 association matrix after generating it (thereby keeping the same relationship between the variance in
500 the network edges and in the kinship matrix). In the second scenario, associations are kin-biased by
501 keeping the kinship matrix as it was generated, thus strongly-bonded associates have the highest
502 kinship on average and weak associates and non-associates (missing edges) have on average the
503 lowest kinship.
504 We simulate 100 replications for combinations of network sizes, with number of individuals
505 ranging from $N = 5$ to $N = 120$ and for the mean numbers of observations per individual ranging
506 from $K = 5$ to $K = 40$. For each run of the simulation, we calculate P values of the effect of the trait
507 value on degree using (1) node permutation tests with the coefficient value as the test statistic, (2)
508 node permutation tests controlling for number of observations, (3) pre-network permutation tests
509 with the coefficient value as the test statistic, (4) pre-network permutation tests with the t statistic as
510 the test statistic, and (5) double permutation tests with the coefficient as the test statistic. We create
511 the networks, conduct the pre-network permutation tests, and conduct the regressions using the
512 MRQAP functionality in the R package *asnipe* (Farine 2013).
513
514
515 **THE DOUBLE PERMUTATION APPROACH IS ROBUST TO TYPE I AND TYPE II ERRORS**
516
517 Our simulations confirm that when there are no effects ($T_S = FALSE$ and $L_S = FALSE$), pre-
518 network permutation tests are prone to elevated false positives (type I error rate of 26%, Figure S1,
519 Table 1), confirming previous studies. Our simulations also show that the tendency for pre-network
520 permutation tests to generate type I errors is greater in smaller networks and when more data are
521 collected (Figure S1). When the t value is used as a test statistic instead of the coefficient, pre-
522 network permutation tests are still prone to false positives (type I error rate of 14%, Table 1), and also
523 perform relatively poorly at detecting a real effect (detecting true effects approximately 20% less
524 often than other approaches, Figure S2, Table 1). The node permutation tests perform particularly
525 poorly when the effects are driven by non-social factors, such as variation in the spatial distribution
526 of individuals (type I error rate of 85%, Figure S3, Table 1). By contrast, double permutation tests
527 perform largely as expected throughout the parameter space, producing conservative P values when
528 no effect is present (type I error rate of 5%), reliably detecting effects when they were present (in line
529 with other tests, Table 1), and being much more conservative than other tests when the effect is
530 driven by non-social factors (type I error rate of 10%, Figures S1-S3, Table 1).
531
532

	No effects ($T_S = FALSE$ and $L_S = FALSE$)	**Social effect ($T_S = TRUE$ and $L_S = FALSE$)**	**Spatial confound ($T_S = TRUE$ and $L_S = TRUE$)**
Node permutation (β)	**4.9%**	**84.8%**	84.9%
Node permutation controlling for number	**5.1%**	**87.3%**	87.8%

of observations (β)			
Pre-network permutation (β)	26.3%	**88.1%**	22.9%
Pre-network permutation (t)	13.6%	62.9%	28.2%
Double permutation	**5.1%**	**86.9%**	**9.9%**

Table 1. Propensity for permutation tests to yield errors or detect real effects when using regression models to test hypotheses on network data (model 1). Table shows the proportion of statistically significant results for an effect of a trait on degree under three sets of scenarios. When $T_S = FALSE$ and $L_S = FALSE$, the expected proportion of significant results should be approximately 5%. When $T_S = TRUE$ and $L_S = FALSE$, the simulated data should have a strong social effect that, and most results should be significant. When both $T_S = TRUE$ and $L_S = TRUE$, the simulated data should have a strong spatial 'nuisance' effect, with the local density of individuals varying across space, and the proportion of significant results should again approach 5%. Figures S1-S3 show how the proportion of significant results is affected by the number of observations and the number of nodes in the network. Bold values highlight test results that performed relatively well.

The problem of elevated error rates is not one of how the data are collected, but rather how the biological inference is drawn from a given dataset. To demonstrate this, we also show the applicability of our combined node permutation solution to data collected from focal observations. Puga-Gonzalez, Sueur and Sosa (2020) recently published the results from simulations (originally based on Farine 2017) showing that the same issue with pre-network permutations using group observations also exists for data collected using focal observations. By simulating scenarios combining both the presence/absence of an effect (females are more social) as well as the presence/absence of a strong observation bias (females are missed 20% of the time), Puga-Gonzalez, Sueur and Sosa (2020) also confirm that node permutations generate substantial rates of type I errors (false positives) when non-social nuisance effects are present.

Using the same simulation code, we show that our double permutation test performs well across all four scenario combinations (Table 2). It is a more conservative approach than pre-network permutation tests alone (remaining close to 5% false positives), performs adequately in terms of type II errors, for example by being less prone to nuisance effects when compared to node permutations.

Note that the true number of real positives is not actually known, and therefore the proportion of type II errors estimated by the simulations is likely to be over-inflated, as not all the simulations will have produced data with an effect present. Finally, because we use this simulation to explore effect size issues (see following section), we report the results of node permutations while controlling for the number of observations there.

	No observation bias		**Observation bias ('nuisance' effect)**	
	Phenotypes equal (Type I errors)	**Females more social (Type II errors)**	**Phenotypes equal (Type I errors)**	**Females more social (Type II errors)**
Node permutation (β)	4.6%	2.4%	57.8%	47.6%
Pre-network permutation (β)	38.2%	10.0%	37.2%	13.2%
Pre-network permutation (t)	28.4%	47.0%	53.4%	44.0%
Double permutation	5%	18.0%	7.2%	24.4%

Table 2. Propensity for permutation tests to produce type I and type II errors from datasets simulating focal sampling (model 2). Simulations use the code from Puga-Gonzalez, Sueur and Sosa (2020), that comprised four scenarios: (1) females and males have identical social phenotypes and are observed equally, (2) females are more social and both sexes are observed equally, (3) females and males have identical social phenotypes but observations are biased towards males (20% of observations of females are missed), and (4) females are more social but observations are biased towards males (20% of observations of females are missed). Using double permutation tests has relatively conservative type I and type II error rates across scenarios. False positives (type I errors) are near 5%, avoiding the high rates suffered by pre-network permutations alone and by node permutations in the presence of nuisance effects.

Finally, we show that the double permutation test is robust to the presence of nonrandom social structure (similar to a node permutation test). A number of social effects can simultaneously shape the structure of social networks (Figure 1), thereby increasing variance in both node-based metrics (e.g. degree) and edge-based metrics (e.g. edge weights). Such high variance can then lead to elevated type I error rates (Weiss *et al.* 2020). Simulations comprised two scenarios, testing the hypothesis that edge weights are predicted by kinship in networks in networks where the association rates (edge weights) are related to kinship and in networks where they are not. As with the above two simulations, the double permutation test performs as expected when no real effect is present (i.e. type I error rates were close to 5%, Table 3). All models have elevated type II error rates because not all simulated networks result in a strong effect present, but the double permutation test performs more conservatively than node permutations (producing more type II errors, Table 3). While pre-network permutations appear to outperform other approaches with respect to Type II errors, this is likely because they are also more sensitive to weak effects in small networks, which are likely to correspond to type I errors rather than correctly identifying a true effect (see Figure S5, which shows higher rates of significant effects in small networks with high numbers of observations).

	Kinship ≠ Associations (Type I errors)	**Kinship ∝ Associations (Type II errors)**
Node permutation (β)	5.1%	16.7%
Pre-network permutation (β)	18.6%	9.9%
Pre-network permutation (t)	3.0%	80.6%
Double permutation	5.2%	22.1%

Table 3. Propensity for permutation tests to produce type I and type II errors regarding kinship effects from simulated datasets with confounding social effects, i.e. nonrandom social structure (model 3). Table shows the type I error rates in simulations where the social effect is a confound (i.e. strong associations are not linked to kinship), and estimated type II error rates in simulations where the social effect corresponds to the hypothesis being tested (i.e. strong associations are linked to kinship). Figures S4-S5 show how the proportion of significant results is affected by the number of observations and the number of nodes in the network.

In summary, our results suggest that double permutation tests are most useful when sampling biases or other nuisance effects might be an issue, and especially when the impact of such effects are expected but not well understood. This method is an alternative to model-fitting methods, such as fitting generalized additive models that can handle non-linearity in the relationship between

sampling intensity and a network metric of interest (Franks *et al.* 2020). Some sampling biases (such as spatial variation in the proportion of individuals that can be identified) are quite complex to model. However, in situations where there is good reason to believe that network data are unbiased, node permutations (or restricted node permutations, see alternative approaches section) can perform well.

THE CHALLENGE OF CALCULATING EFFECT SIZES

Inference will always benefit from relying less on *P* values and instead focusing more on effect sizes (Nakagawa & Cuthill 2007). Franks *et al.* (2020) proposed that the coefficients of models can generate reliable relative effect sizes after controlling for the number of observations. However, multiple other nuisance effects can also create problems for estimating effect sizes and their significance. We explored this using the simulation of scenarios in which females are more social but also less observable (using Model 2). While the original coefficient (before the observation bias) and estimated coefficient (with the observation bias) were correlated (r=0.54), controlling for the number of observations of each individual consistently inflated the estimated coefficient size (Figure S6). We tested whether regression models can recover the original coefficient value using two approaches to fitting the number of observations as a covariate. First, we used a naïve model, whereby the scaled number of observations is simply added as a covariate. Second, we used a more informed model whereby the number of observations is added as an interaction with the effect of sex (exploration of the data would show that the number of observations differs between sexes). The naïve model performed worse, producing estimated effect sizes that were on average 1.8 times the original value (and up to 5.1 times the original value). Correctly fitting observations as an interaction term did not dramatically improve this, with the average estimated coefficient values being 1.7 times the original value (and up to 3.3 times the original). These two models performed even worse at estimating effect sizes when the true effect was not present (the estimated effect sizes were on average over 250 times the true values, Figure S7).

One reason why the models could not generate robust effect sizes is because the models do not deal well with correcting data in situations where individuals are observed in groups rather than in pairs. For group observations, the loss of each observation can result in a variable number of edges being removed, with variation occurring both within groups (missing one individual from a group of 10 will reduce its degree by 9 units whereas others' degree will only reduce by 1 unit) and between groups (missing one individual will result in a larger loss of edges in a larger group than in a small group). Given our findings, approaches to estimating corrected effect sizes should be carefully tested before being used. Estimating effect sizes in the presence of bias is a major priority in the continued development of robust tools for animal social network analysis.

ALTERNATIVE APPROACHES

While our double permutation test performs similarly, and generally better, than the single permutation procedures across a range of scenarios, many alternative approaches or methodological refinements can improve the robustness of inferences from hypothesis testing. Here we discuss some alternative and/or further approaches.

For many studies, it may not be necessary to use a double permutation test. It will often be sufficient to use node permutations and control for nuisance effects by restricting which individuals'

data are swapped when performing the randomization. Such restricted node permutations are useful if individuals can be easily allocated to a distinct spatial location, or if there are clear categories of individuals that correspond to biases. For example, if individual animals enter the study in distinct waves, because of a standard dispersal time or because a study expanded at some point to include new individuals, then node permutations could be restricted among individuals entering the study at approximately the same time. However, if multiple factors have to be accounted for, one can rapidly run out of sets of individuals to swap. For example, a study population comprising 40 individuals that aims to restrict swaps by two parameters (e.g. age and location) would have on average only 10 individuals per class if these are binary, only 6-7 per class if one is trinary, and only 4-5 per class if both are trinary.

Another approach is to explicitly estimate the uncertainty of the combination of a given dataset and hypothesis-testing method. The procedure we used here—simulating a random trait value for each node in a network and running through the full hypothesis-testing procedure—can be a straightforward way of characterising the robustness of any given study's results. That is, one can explore how sensitive a given dataset is to generating false positives or false negatives under different hypothesis-testing approaches. This procedure simply involves generating a random trait variable (e.g. drawing a trait value from a normal distribution) and testing how this value corresponds to the metric of interest from the observed network using the same code as for the real variable(s) being studied. By repeating this procedure many times, the proportion of the tests that produce a significant P value can be reported. It is worth exploring if and how this study-specific information might be used. For example, one might be able to correct the threshold for rejecting the null hypothesis to the point where the expected false positive rate will be 5%. Shuffling the actual node values (even doing so within space or time) and repeatedly running pre-network permutation tests might provide an even more precise estimation of the true false positive rate, as it will be fully conditioned on the real observation data.

Rather than testing one null hypothesis, which encourages confirmation bias, the principle of strong inference (Platt 1964) requires considering and testing multiple alternative hypotheses. One criticism of null hypothesis testing using permutation tests is that they do not provide an exhaustive exclusion of alternative explanations (Zhang 2020). However, this apparent weakness can be turned into a strength if multiple models are used to collectively examine the different processes that might be shaping the patterns present in observation data. The use of multiple permutation-based null models can therefore be highly informative. For example, while it is important to control for the contribution of 'nuisance' spatial effects to social network structure when testing hypotheses about social decision-making, how animals use space (and its links to social structure) is itself an important biological process (Webber & Vander Wal 2018; He, Maldonado-Chaparro & Farine 2019). We show an example of this in Figure 3, where both social and spatial processes shape the differences in the group sizes of males and females, and where pre-network permutations that control for space would discard the biological drivers of space use (and, consequently, group size) as a nuisance effect. Aplin *et al.* (2015) evaluated the extent that the spatial distribution of individuals contributed towards their repeatability in social network metrics by reporting the distribution of repeatability values from a spatially-constrained permutation test. Farine *et al.* (2015) used two different permutation tests to identify the expected effects of individuals choosing social groups versus choosing habitats. Such an approach can characterise the relative drivers of apparent gregariousness among males and females in Figure 3. Hobson, Monster and Dedeo (2019) permuted observations of dominance interactions and then used the direction of the deviations from null expectations for inferring the likelihood of

different dominance strategies, such as individuals that preferentially attacked groupmates that were closer in rank (targeting close competitors) or farther in rank (bullying weaker opponents).

Another approach that is often considered to be useful for estimating uncertainty (e.g. confidence interval around effect sizes) is bootstrapping (Lusseau, Whitehead & Gero 2008; Farine & Strandburg-Peshkin 2015; Bonnell & Vilette 2020). Bootstrapping involves resampling the observed data with replacement to create new datasets of the same size as the original. This procedure can estimate the range of values that a given statistic can take, and whether the estimate overlaps with an expected null value (see Puth, Neuhauser & Ruxton 2015). Bootstrapping, however, is not always appropriate as a means of hypothesis testing in animal social networks, because like node permutations, it relies on resampling the observed data, under the assumption that the observed network reflects the true social structure. For example, missing edges in an association network represent an association rate of zero, but in reality these zero values could be weak associations that exist in the real world, but were simply never observed. Bootstrapping the edge weights incorrectly suggests that unobserved edges have no uncertainty, which is obviously false for most studies. Thus, bootstrapping social network data should only be used with care and for specific aims.

Several other methods that do not rely on permutation tests have been proposed to deal with sampling biases when constructing networks or to deal with non-independence of data to test hypotheses. For example, Gimenez *et al.* (2019) propose using capture-recapture models to explicitly model heterogeneity in detections, thereby providing more accurate estimates of network metrics. Studies estimating phenotypic variance using animal models have also proposed methods to decompose multiple sources driving between-individual variation in trait values (Thomson *et al.* 2018). Such multi-matrix models have recently been applied to animal social networks as a means of identifying the relative importance of different predictors in driving differences in social network metrics (Albery *et al.* 2020).

Pre-network permutation tests were initially designed to evaluate whether the social structure of a population is non-random, given sparse association data (Bejder, Fletcher & Brager 1998). However, as shown by recent studies (Weiss et al. 2020; Puga-Gonzalez et al. 2020), and our own, producing reliable tests of some hypotheses can predictably degrade as the size of the observational dataset increases (Figure S1, for the reasons outlined in Figure 2). With rich datasets, however, it becomes possible to create replicated networks (Hobson, Avery, & Wright, 2013). That is, the data can be split to produce multiple networks (without overlapping observations), where each network contains sufficient data to produce reliable estimates of network structure (Farine, 2018). The same hypothesis-testing procedure can then be applied to each network independently. Using emerging methods for automated tracking, social networks can be created for each season (e.g. Papageorgiou *et al.* 2019), across periods of several days (e.g. Dakin *et al.* 2019), for each day (e.g. Boogert, Farine & Spencer 2014), or right down to each second (e.g. Blonder & Dornhaus 2011). If these networks produce consistent results when tested independently, this provides much stronger support for a given hypothesis than any single network. Alternatively, waxing and waning of effects might suggest some underlying dynamics are present that warrant further investigation, more careful analyses, or longer periods for each replication. Any inference becomes stronger again if each of the replicate social networks contains different sets of individuals, and the networks are re-formed in each sample period, as for instance when networks represent spatial proximity in roosts after individuals come back together to sleep or rest after having foraged or moved individually (Ripperger *et al.* 2019). Given sufficient data, many networks could be combined by using tools from meta-analyses to estimate an overall effect size. However, such an approach would need to ensure that the same biases don't impact each of the networks in the same way.

Although within-study replication can improve our confidence in a given result, ultimately the gold standard is replication across studies. Within-study replications cannot control for many of the nuances in how data are collected, stored, and analysed. One example of a replication study tested the effects of developmental conditions on the social network position of juvenile zebra finches (*Taeniopygia guttata*). In the original study (Boogert, Farine & Spencer 2014), birds were given either stress hormone or control treatments as nestlings, and their social relationships were studied after they became nutritionally independent from their parents. In the replication study (Brandl *et al.* 2019), clutch sizes of wild zebra finches were manipulated to experimentally increase or decrease sibling competition (a source of developmental stress), and social associations (in the wild) were recorded after birds fledged. Across both studies, 9 out of 10 hypothesized effects had the same result (i.e. both statistically significant or not), and all 10 of the differences were in the same direction (binomial $P<0.0001$).

POTENTIAL IMPROVEMENTS

There are many further directions that can be explored for the double-permutation test. While we have demonstrated that our method performs adequately across a range of scenarios using simulated data, it is not always possible to simulate all possible types of uncertainty that might exist in empirical datasets. For example, the median expected value might not effectively represent the value expected by the null hypothesis, because when the possible configurations of the data are severely limited (for instance by a small sample size of observations per individual), the resulting distribution of expected metric values might not be unimodal. For example, if individuals have strong group size preferences, then their expected degree might jump dramatically from their preferred values to the distribution of mean degrees from the population, as more and more swaps are made. It can therefore be useful to visualize the expected distributions of metric values across individuals when possible, and to remove individuals that have been under-sampled (Farine & Strandburg-Peshkin 2015).

Rather than using a single value (such as the median, see Figure 4), future studies could explore ways of carrying over uncertainty from the distribution of permutation values when calculating the corrected test statistic. For example, one could use a Monte Carlo approach that repeatedly samples from the distribution of permuted values when calculating the residual for each unit in the analysis (each node or each edge), and using these many measurements to estimate the 95% range of the corrected test statistic. An alternative way of carrying uncertainty through the analysis could be to implement a Bayesian framework for inferring the network (Farine & Strandburg-Peshkin 2015), or even to modelling the dynamic process of the observations of connections in the network (Koskinen & Snijders 2007). Thus, there remains significant grounds for continued improvements in the methods for conduction hypothesis testing using animal social network data.

For many studies, it is important to not only test a hypothesis of interest but also to accurately estimate the connection strength between individuals. One method that has been proposed are generalized affiliation indices (Whitehead & James 2015). These involve regressing the observed association strength against nuisance factors (such as home-range overlap) to generate a corrected value that accounts for the opportunity to associate. Permutation tests have also been suggested as a means of identifying non-random preferred or avoided relationships (Whitehead, Bejder & Ottensmeyer 2005). Yet, it remains to be determined whether permutation tests could also provide more accurate estimates of the strength of each relationship. Following our methods, it could

be possible to estimate a corrected association (or interaction) strength by subtracting some measure of the distribution of permuted values from the observed value of each edge.

CONCLUSIONS

In this paper, we have revisited some of the factors that can raise problems when conducting hypothesis testing using animal social network data, highlighting the need for more robust methods. By combining the strengths of two randomisation routines, we developed an approach that does not suffer from elevated false positives and suffers relatively little from false negatives. However, our proposed solution, or the use of permutation tests more generally, does not negate the need to carefully consider statistical issues that have been highlighted for more orthodox statistical practices (Forstmeier, Wagenmakers & Parker 2017). For example, the common practice of using linear models for both data exploration and hypothesis testing are estimated to produce rates of type I error as high as 40% (Forstmeier & Schielzeth 2011). High false discovery rates can also be caused by choosing an incorrect model structure (e.g. by failing to fit random slopes to a mixed effects model, see Schielzeth & Forstmeier 2009), which could also be applicable to the method used when calculating a test statistic for use with a permutation test. In general, false positive rates are likely to increase with the complexity of the question and the dataset, and dealing with empirical datasets in the biological sciences often requires making complex decisions for which the solutions aren't clear—such as whether to log-transform data (Ives 2015) or not (O'Hara & Kotze 2010). In the context of social network data, different permutation procedures (including constricting the same test in different ways) each test quite a specific null hypothesis (see Figure 3 for an example), so part of statistical considerations should include ensuring that the correct null hypothesis is being tested.

One particularly important point that our work, and that of others (e.g. Franks *et al.* 2020), highlights is the need to pay particularly close attention to the importance of different processes for a given hypothesis of interest. Take, for example, variation in space use and corresponding differences in the local density of individuals. If we assume that space use is constrained by non-social factors, and if we aim to understand animal social decisions, then space use (and its consequences for density) could be considered a nuisance effect. However, if we aim to study the transmission of information or pathogens, then these same effects are now an important factors contributing to the outcome of the transmission process. Thus, a given factor might represent a nuisance effect for one question but not another, even if these factors represent two halves of the same feedback loop (Cantor *et al.* 2019). Unfortunately, a major challenge remains in estimating the importance of effect sizes, in the presence of nuisance effects, when using animal social network data.

Using three different simulations, each including multiple scenarios, we have shown that pre-network permutation tests can produce reliable results when combined with node permutations to form a double permutation test. In contrast to parametric approaches (or drawing heavily from these when producing a test statistic), they can control for a large range of nuisance effects without implementing complex model structures, measuring and controlling for every source of bias, or assessing the consequences of deviating from parametric model assumptions. Further, they avoid making assumptions that the observed network corresponds exactly to the true network. Using permutation tests requires and encourages researchers to focus on thinking carefully about what specific processes may have produced the patterns in a given observed dataset. One can use a range of permutation tests to evaluate the relative contribution of different processes by measuring the

relative deviations of the data away from their null expectations. The strength and robustness of permutation tests therefore lies in their flexibility and simplicity.

ACKNOWLEDGEMENTS

We thank the Farine lab, Josh Firth, Matt Silk, Michael Weiss, Dan Franks, and the many researchers who contacted the authors with questions, for helpful comments, insights, and discussions regarding the issues presented in this paper. DRF was funded by the Max Planck Society and a grant from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No. 850859). DRF received additional from the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – EXC 2117 – 422037984.

CODE AVAILABILITY

Code for simulation models 1 to 3 is available here:
https://owncloud.gwdg.de/index.php/s/93zgi9tiKlu5Grr

REFERENCES

Alarcón-Nieto, G., Graving, J.M., Klarevas-Irby, J.A., Maldonado-Chaparro, A.A., Mueller, I. & Farine, D.R. (2018) An automated barcode tracking system for behavioural studies in birds. *Methods in Ecology and Evolution,* **9,** 1536-1547.

Albery, G.F., Morris, A., Morris, S., Pemberton, J.M., Clutton-Brock, T.H., Nussey, D.H. & Firth, J.A. (2020) Spatial point locations explain a range of social network positions in a wild ungulate. *bioRxiv*, https://doi.org/10.1101/2020.1106.1104.135467.

Aplin, L.M., Firth, J.A., Farine, D.R., Voelkl, B., Crates, R.A., Culina, A., Garroway, C.J., Hinde, C.A., Kidd, L.R., Psorakis, I., Milligan, N.D., Radersma, R., Verhelst, B. & Sheldon, B.C. (2015) Consistent individual differences in the social phenotypes of wild great tits (*Parus major*). *Animal Behaviour,* **108,** 117-127.

Bejder, L., Fletcher, D. & Brager, S. (1998) A method for testing association patterns of social animals. *Animal Behaviour,* **56,** 719-725.

Blonder, B. & Dornhaus, A. (2011) Time-Ordered Networks Reveal Limitations to Information Flow in Ant Colonies. *Plos One,* **6**.

Bonnell, T.R. & Vilette, C. (2020) Constructing and analysing time-aggregated networks: The role of bootstrapping, permutation and simulation. *Methods in Ecology and Evolution,* **In press**.

Boogert, N.J., Farine, D.R. & Spencer, K.A. (2014) Developmental stress predicts social network position. *Biology Letters,* **10,** 20140561.

Brandl, H.B., Farine, D.R., Funghi, C., Schuett, W. & Griffith, S.C. (2019) Early-life social environment predicts social network position in wild zebra finches. *Proceedings of the Royal Society B-Biological Sciences,* **286**.

Cantor, M., Maldonado-Chaparro, A., Beck, K., Carter, G.G., He, P., Hilleman, F., Klarevas-Irby, J.A., Lang, S.D.J., Ogino, M., Papageorgiou, D., Prox, L. & Farine, D.R. (2019) Animal social networks: revealing the causes and implications of social structure in ecology and evolution. *EcoEvoRxiv,* **https://doi.org/10.32942/osf.io/m62gb**.

Crall, J.D., Gravish, N., Mountcastle, A.M. & Combes, S.A. (2015) BEEtag: A Low-Cost, Image-Based Tracking System for the Study of Animal Behavior and Locomotion. *Plos One,* **10,** e0136487.

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

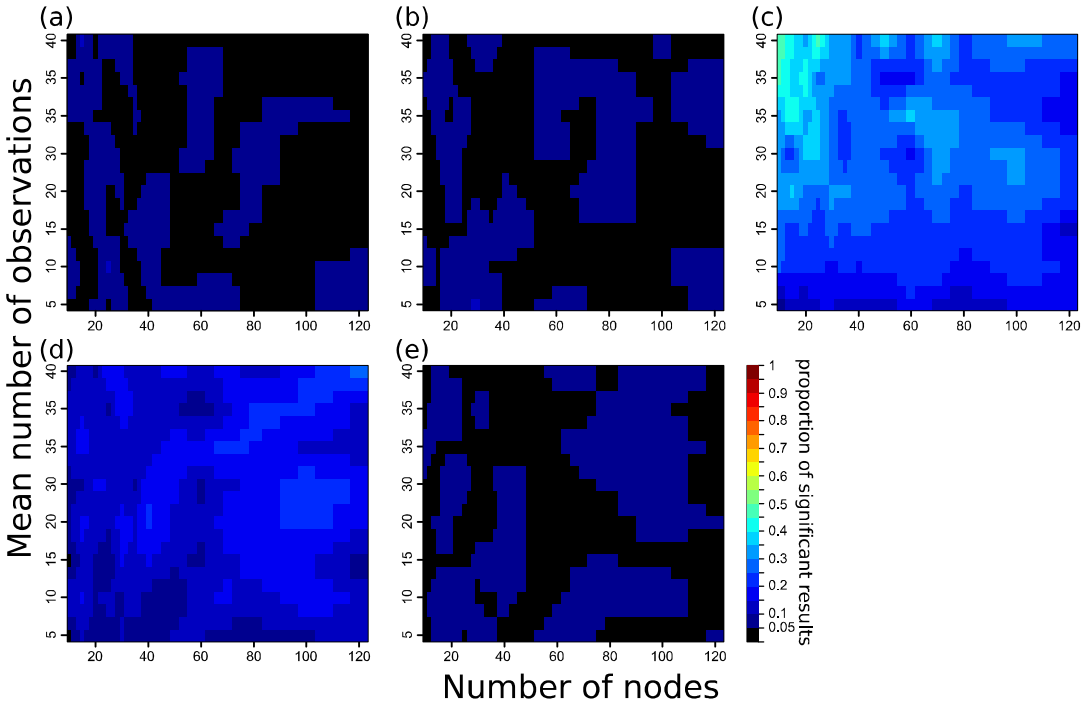
Dakin, R., Moore, I.T., Horton, B.M., Vernasco, B.J. & Ryder, T.B. (2019) Testosterone-mediated behavior shapes the emergent properties of social networks. *bioRxiv,* **10.1101/737650**.

Davis, G.H., Crofoot, M.C. & Farine, D.R. (2018) Estimating the robustness and uncertainty of animal social networks using different observer methods. *Animal Behaviour,* **141,** 29-44.

884 Farine, D.R. (2013) Animal Social Network Inference and Permutations for Ecologists in R using asnipe. *Methods*
885 *in Ecology and Evolution,* **4,** 1187–1194.
886 Farine, D.R. (2017) A guide to null models for animal social network analysis. *Methods in Ecology and Evolution,*
887 **8,** 1309-1320.
888 Farine, D.R. & Aplin, L.M. (2019) Spurious inference when comparing networks. *Proceedings of the National*
889 *Academy of Sciences of the United States of America,* **116,** 16674-16675.
890 Farine, D.R., Firth, J.A., Aplin, L.M., Crates, R.A., Culina, A., Garroway, C.J., Hinde, C.A., Kidd, L.R., Milligan, N.D.,
891 Psorakis, I., Radersma, R., Verhelst, B., Voelkl, B. & Sheldon, B.C. (2015) The role of social and
892 ecological processes in structuring animal populations: a case study from automated tracking of wild
893 birds. *Royal Society Open Science,* **2,** 150057.
894 Farine, D.R. & Strandburg-Peshkin, A. (2015) Estimating uncertainty and reliability of social network data using
895 Bayesian inference. *Royal Society Open Science,* **2,** 150367.
896 Farine, D.R. & Whitehead, H. (2015) Constructing, conducting, and interpreting animal social network analysis.
897 *Journal of Animal Ecology,* **84,** 1144-1163.
898 Firth, J.A., Cole, E.F., Ioannou, C.C., Quinn, J.L., Aplin, L.M., Culina, A., McMahon, K. & Sheldon, B.C. (2018)
899 Personality shapes pair bonding in a wild bird social system. *Nature Ecology & Evolution,* **2,** 1696-1699.
900 Forstmeier, W. & Schielzeth, H. (2011) Cryptic multiple hypotheses testing in linear models: overestimated
901 effect sizes and the winner's curse. *Behavioral Ecology and Sociobiology,* **65,** 47-55.
902 Forstmeier, W., Wagenmakers, E.J. & Parker, T.H. (2017) Detecting and avoiding likely false-positive findings - a
903 practical guide. *Biological Reviews,* **92,** 1941-1968.
904 Franks, D.W., Weiss, M.N., Silk, M.J., Perryman, R.J.Y. & Croft, D.P. (2020) Calculating effect sizes in animal social
905 network analysis. *Methods in Ecology and Evolution,* **10.1111/2041-210X.13429**.
906 Gimenez, O., Mansilla, L., Klaich, M.J., Coscarella, M.A., Pedraza, S.N. & Crespo, E.A. (2019) Inferring animal
907 social networks with imperfect detection. *Ecological Modelling,* **401,** 69-74.
908 He, P., Maldonado-Chaparro, A. & Farine, D.R. (2019) The role of habitat configuration in shaping social
909 structure: a gap in studies of animal social complexity. *Behavioral Ecology and Sociobiology,* 9.
910 Hobson, E.A., Monster, D. & Dedeo, S. (2019) Strategic heuristics underlie animal dominance hierarchies and
911 provide evidence of group-level social knowledge. *arXiv*.
912 Ives, A.R. (2015) For testing the significance of regression coefficients, go ahead and log-transform count data.
913 *Methods in Ecology and Evolution,* **6,** 828-835.
914 Koskinen, J.H. & Snijders, T.A.B. (2007) Bayesian inference for dynamic social network data. *Journal of Statistical*
915 *Planning and Inference,* **137,** 3930-3938.
916 Langen, T.A. (1996) Social learning of a novel foraging skill by white-throated magpie-jays (Calocitta formosa,
917 Corvidae): A field experiment. *Ethology,* **102,** 157-166.
918 Lantz, B. (2013) The large sample size fallacy. *Scandinavian Journal of Caring Sciences,* **27,** 487-492.
919 Leedale, A.E., Sharp, S.P., Simeoni, M., Robinson, E.J.H. & Hatchwell, B. (2018) Fine-scale genetic structure and
920 helping decisions in a cooperatively breeding bird. *Molecular Ecology,* **27,** 1714-1726.
921 Lusseau, D., Whitehead, H. & Gero, S. (2008) Incorporating uncertainty into the study of animal social networks.
922 *Animal Behaviour,* **75,** 1809-1815.
923 Nakagawa, S. & Cuthill, I.C. (2007) Effect size, confidence interval and statistical significance: a practical guide
924 for biologists. *Biological Reviews,* **82,** 591-605.
925 O'Hara, R.B. & Kotze, D.J. (2010) Do not log-transform count data. *Methods in Ecology and Evolution,* **1,** 118-
926 122.
927 Papageorgiou, D., Christensen, C., Gall, G.E., Klarevas-Irby, J.A., Nyaguthii, B., Couzin, I.D. & Farine, D.R. (2019)
928 The multilevel society of a small-brained bird. *Current Biology,* **29,** R1120-R1121.
929 Platt, J.R. (1964) Strong Inference. *Science,* **146,** 347-353.
930 Puga-Gonzalez, I., Sueur, C. & Sosa, S. (2020) Null models for animal social network analysis and data collected
931 via focal sampling: Pre-network or node network permutation? *Methods in Ecology and Evolution,*
932 **10.1111/2041-210X.13400**.
933 Puth, M.T., Neuhauser, M. & Ruxton, G.D. (2015) On the variety of methods for calculating confidence intervals
934 by bootstrapping. *Journal of Animal Ecology,* **84,** 892-897.
935 Ripperger, S.P., Carter, G.G., Duda, N., Koelpin, A., Cassens, B., Kapitza, R., Josic, D., Berrio-Martinez, J., Page,
936 R.A. & Mayer, F. (2019) Vampire Bats that Cooperate in the Lab Maintain Their Social Networks in the
937 Wild. *Current Biology,* **29,** 4139-+.
938 Ripperger, S.P., Carter, G.G., Page, R.A., Duda, N., Koelpin, A., Weigel, R., Hartmann, M., Nowak, T., Thielecke, J.,
939 Schadhauser, M., Robert, J., Herbst, S., Meyer-Wegener, K., Wagemann, P., Schroder-Preikschat, W.,
940 Cassens, B., Kapitza, R., Dressler, F. & Mayer, F. (2020) Thinking small: Next-generation sensor networks
941 close the size gap in vertebrate biologging. *Plos Biology,* **18**.

942 Ryder, T.B., Horton, B.M., van den Tillaart, M., Morales, J.D. & Moore, I.T. (2012) Proximity data-loggers increase
943 the quantity and quality of social network data. *Biology Letters,* **8,** 917-920.
944 Schielzeth, H. & Forstmeier, W. (2009) Conclusions beyond support: overconfident estimates in mixed models.
945 *Behavioral Ecology,* **20,** 416-420.
946 Szucs, D. & Ioannidis, J.P.A. (2017) When Null Hypothesis Significance Testing Is Unsuitable for Research: A
947 Reassessment. *Frontiers in Human Neuroscience,* **11**.
948 Thomson, C.E., Winney, I.S., Salles, O.C. & Pujol, B. (2018) A guide to using a multiple-matrix animal model to
949 disentangle genetic and nongenetic causes of phenotypic variance. *Plos One,* **13**.
950 Webber, Q.M.R. & Vander Wal, E. (2018) An evolutionary framework outlining the integration of individual
951 social and spatial ecology. *Journal of Animal Ecology,* **87,** 113-127.
952 Weiss, M.N., Franks, D.W., Brent, L.J.N., Ellis, S., Silk, M.J. & Croft, D.P. (2020) Common datastream permutations
953 of animal social network data are not appropriate for hypothesis testing using regression models.
954 *bioRxiv,* **10.1101/2020.04.29.068056**.
955 Whitehead, H. (2008) *Analyzing animal societies.* University of Chicago Press, Chicago, USA.
956 Whitehead, H., Bejder, L. & Ottensmeyer, C.A. (2005) Testing association patterns: issues arising and extensions.
957 *Animal Behaviour,* **69,** e1-e6.
958 Whitehead, H. & James, R. (2015) Generalized affiliation indices extract affiliations from social network data.
959 *Methods in Ecology and Evolution,* **6,** 836-844.
960 Zhang, M.J. (2020) The use and limitations of null-model-based hypothesis testing. *Biology & Philosophy,* **35**.
961
962
963

964 **APPENDIX**
965



966
967
968 **Figure S1. Proportions of statistically significant results for differently-sized networks and different observation efforts for**
969 **simulation 1 under the scenario when $T_S = FALSE$ (no effect is present) and $L_S = FALSE$ (individuals are not**
970 **preferentially located in different patches).** Panels represent the P values calculated using (a) node permutation tests with
971 the coefficient value as the test statistic, (b) node permutation tests controlling for number of observations, (c) pre-network
972 permutation tests with the coefficient value as the test statistic, (d) pre-network permutation tests with the t statistic as the
973 test statistic, and (e) double permutation tests with the coefficient as the test statistic. These results highlight the propensity
974 for pre-network permutation tests (c) to produce spurious results when networks have few nodes but many observations
975 (top left of the plot).
976
977

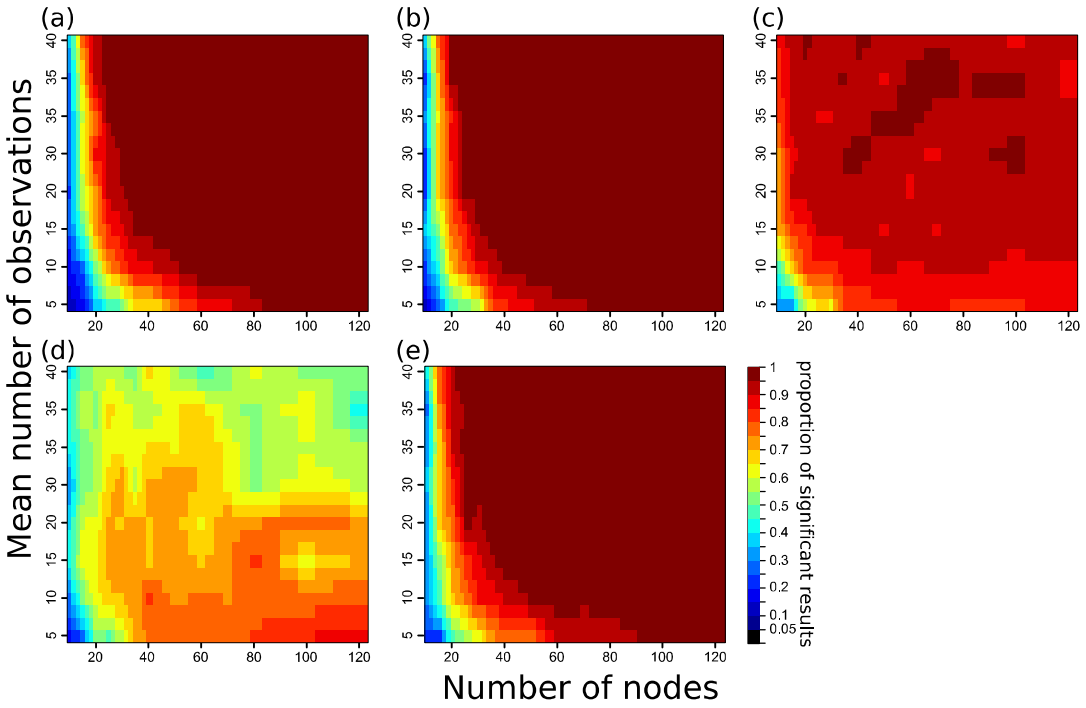


Figure S2. Proportions of statistically significant results for differently-sized networks and different observation efforts for simulation 1 under the scenario when $T_S = TRUE$ (an effect is present) and $L_S = FALSE$ (the effect is not a spatial confound). Panels represent the P values calculated using (a) node permutation tests with the coefficient value as the test statistic, (b) node permutation tests controlling for number of observations, (c) pre-network permutation tests with the coefficient value as the test statistic, (d) pre-network permutation tests with the t statistic as the test statistic, and (e) double permutation tests with the coefficient as the test statistic. These results highlight the propensity for pre-network permutation tests (c) to be more likely to produce significant results when networks have few nodes but many observations (left-hand of the plot relative to panels a, b and e). Further, results show that using the t statistic (d) produces unreliable results (i.e. the significance does not increase when more observations are made).

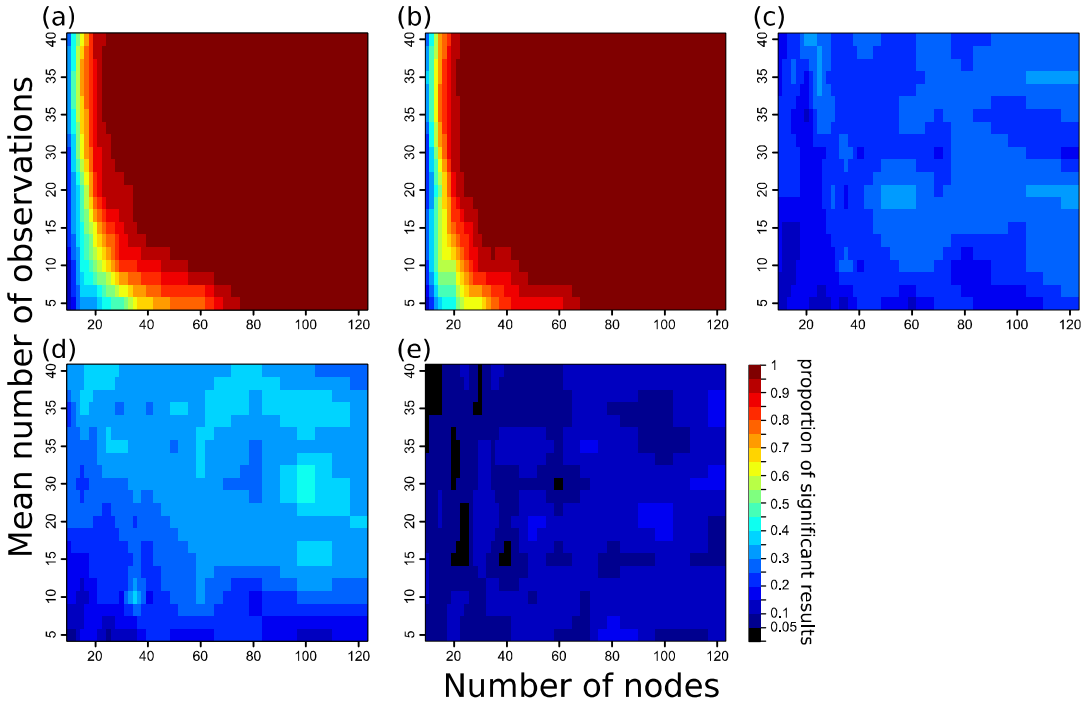


Figure S3. Proportions of statistically significant results for differently-sized networks and different observation efforts for simulation 1 under the scenario when $T_S = TRUE$ (an effect is present) and $L_S = TRUE$ (the effect is a spatial confound). Panels represent the P values calculated using (a) node permutation tests with the coefficient value as the test statistic, (b) node permutation tests controlling for number of observations, (c) pre-network permutation tests with the coefficient value as the test statistic, (d) pre-network permutation tests with the t statistic as the test statistic, and (e) double permutation tests with the coefficient as the test statistic. These results highlight the poor performance of node permutation-based models (panels a and b).

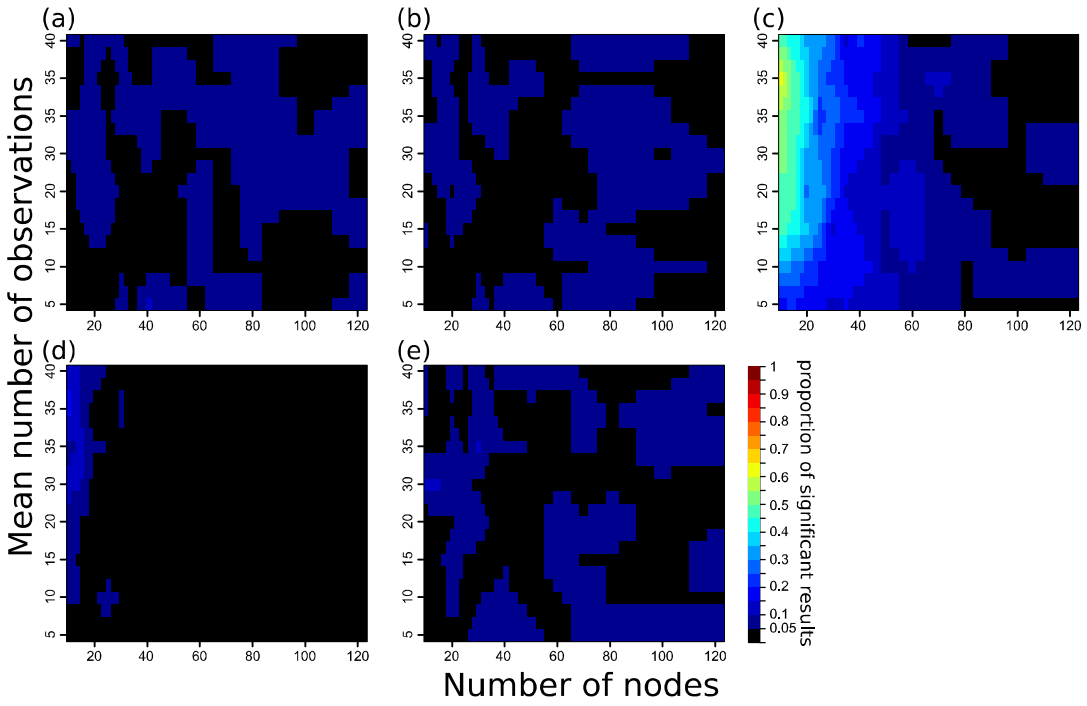


Figure S4. Proportions of statistically significant results for differently-sized networks and different observation efforts for simulation 3 under the scenario where kinship does not predict association rates (edge weights). Panels represent the P values calculated using (a) node permutation tests with the coefficient value as the test statistic, (b) node permutation tests controlling for number of observations, (c) pre-network permutation tests with the coefficient value as the test statistic, (d) pre-network permutation tests with the t statistic as the test statistic, and (e) double permutation tests with the coefficient as the test statistic. These results highlight the propensity for pre-network permutation tests (c) to produce spurious results when networks have few nodes present (left-hand-side of the plot).

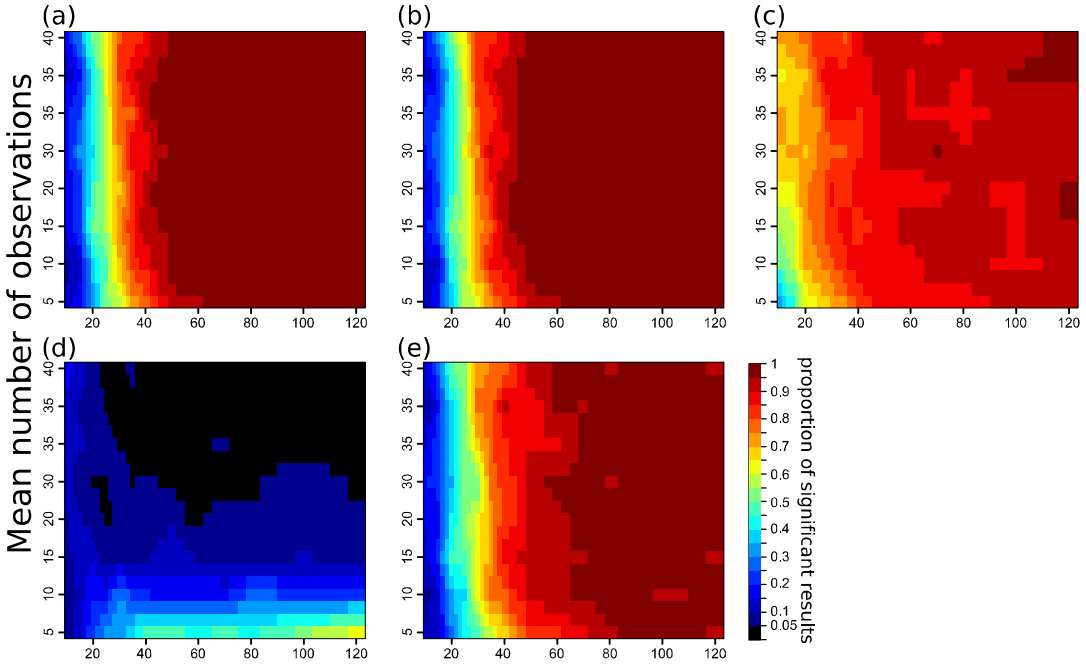


Figure S5. Proportions of statistically significant results for differently-sized networks and different observation efforts for simulation 3 under the scenario where kinship predicts association rates (edge weights). Panels represent the P values calculated using (a) node permutation tests with the coefficient value as the test statistic, (b) node permutation tests controlling for number of observations, (c) pre-network permutation tests with the coefficient value as the test statistic, (d) pre-network permutation tests with the t statistic as the test statistic, and (e) double permutation tests with the coefficient as the test statistic. These results highlight the propensity for pre-network permutation tests (c) to be more likely to produce significant results when networks have few nodes but many observations (left-hand of the plot relative to panels a, b and e). Further, results show that using the t statistic (d) produces unreliable results (i.e. the significance does not increase with when more observations are made).

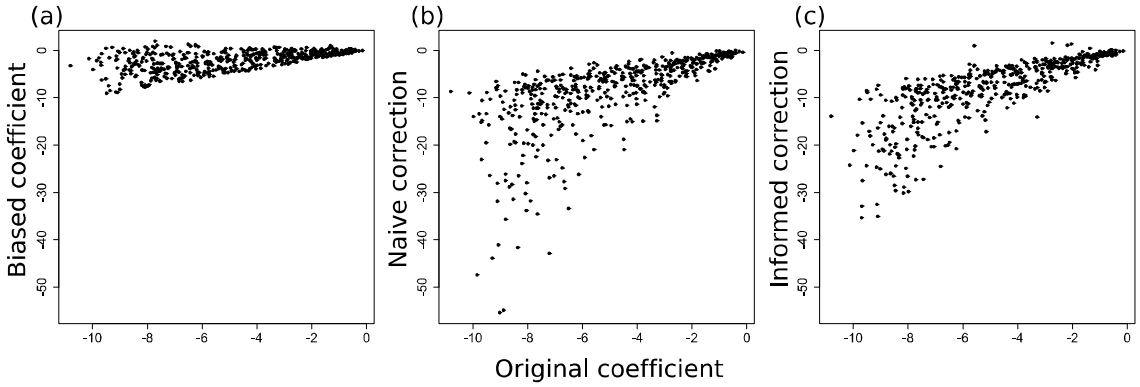


Figure S6. Relationship between the original coefficient value (the relationship between degree and sex prior to introducing an observation bias) and estimations of the coefficient value using the data from simulation 2 where an effect is present (females are more gregarious). (a) The original coefficient versus the coefficient estimated from the biased observations. (b) The original coefficient versus a naïve correction involving adding only the number of observation for each individual as a covariate in the model. (c) The original coefficient versus an informed correction that involves including an interaction term between sex and the number of observations. Each point represents one simulation. While these coefficients are correlated, the corrected coefficient values can be greatly over-estimated, suggesting that adding the number of observations into a model does not produce reliable effect sizes.

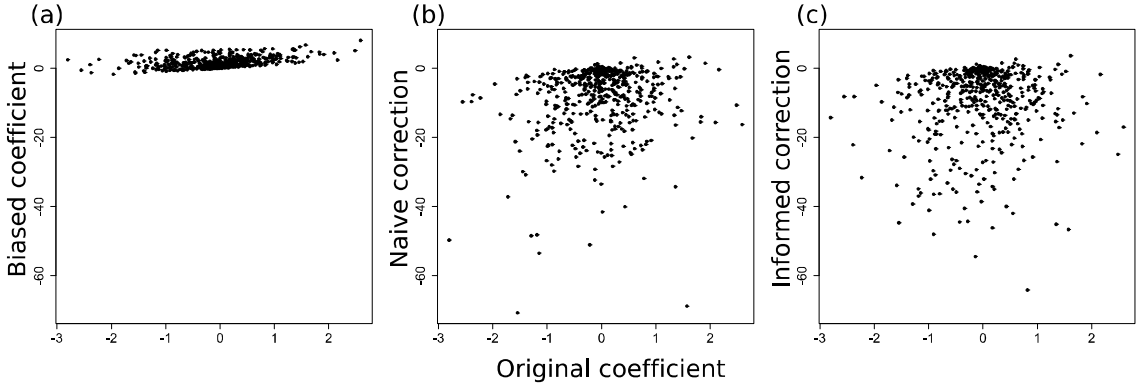


Figure S7. Relationship between the original coefficient value (the relationship between degree and sex prior to introducing an observation bias) and estimations of the coefficient value using the data from simulation 2 where no effect is present (females and males are equally gregarious). (a) The original coefficient versus the coefficient estimated from the biased observations. (b) The original coefficient versus a naïve correction involving adding only the number of observation for each individual as a covariate in the model. (c) The original coefficient versus an informed correction that involves including an interaction term between sex and the number of observations. Each point represents one simulation. Because there was no original effect present, the coefficients are not correlated. However, the corrected coefficient values generate extremely large coefficient values, suggesting that adding the number of observations into a model does not produce reliable effect sizes.