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

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

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

Correspondence: dfarine@ab.mpg.de , *Phone: +49 7732 15054*

Running title: Reliable inference from animal social network data

Statement of Authorship: DRF developed the idea, DRF and GGC outlined the project, DRF conducted the simulations, DRF and GGC wrote the paper.

Data accessibility statement: This submission contains no primary data. The R function and code to replicate the simulations are available to reviewers here: https://owncloud.gwdg.de/index.php/s/93zgi9tiKlu5Grr and will be uploaded to a repository upon acceptance.

Keywords: animal social networks, hypothesis testing, permutation tests, social behaviour, social network analysis

ABSTRACT

1. Permutation tests are widely used to test hypotheses with animal social network data, but suffer from high rates of type I and II error when the permutations do not properly simulate the null hypothesis of interest.
2. Two common types of permutations each have limitations. Pre-network (datastream) permutations can be used to control "nuisance effects" like spatial, temporal, or sampling biases, but only when the null hypothesis assumes random social decisions. Node (node-label) permutation tests can test null hypotheses that include nonrandom social structure, but only when the observed network is not shaped by nuisance effects.
3. We propose a new procedure, the double permutation test, that addresses these limitations by combining pre-network and node permutations. We conduct a range of simulations, allowing us to estimate error rates caused by confounding effects of social or non-social structure in the raw data.
4. We show that the double permutation test is less likely to produce elevated error rates relative to node permutations, pre-network permutations, and regression models with control variables, which all exhibit elevated type I errors under at least one set of simulated conditions. In scenarios where type I error rates from pre-network permutation tests exceed 30%, the error rates from double permutation remain close to 5%.
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 for more robust inferences, including controlled node permutations, testing multiple null hypotheses, and splitting large datasets to generate replicated networks. Finally, we highlight ways that uncertainty can be explicitly considered and carried through the analysis.

INTRODUCTION

Permutation tests are arguably 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 *et al.* 2009) and social networks (Croft *et al.* 2011). Permutation tests randomize (or re-assign) observed data to generate a distribution of statistic values expected under a given null hypothesisresearchers 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). They are particularly useful when the standard assumptions of other statistical tests are violated, as is often the case with social network data (see Farine 2017 for a general introduction). To understand how social network structure differs from what is expected if animals made random social decisions, a researcher can permute the observed data to create many expected networks that could have occurred in the absence of social preferences (*pre-network permutations*). Alternatively, a researcher can preserve the same observed network properties in all expected networks and only permute the node labels (*node permutations*). The difference in the design of these two permutation approaches has important consequences for our ability to make reliable inferences, because they assess different null hypotheses.

Several recent studies (Evans, Fisher & Silk 2020; Puga-Gonzalez, Sueur & Sosa 2021; Weiss *et al.* 2021) 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 is generated using social network data. Consider the question: Does kinship influence association rate? Testing the null hypothesis here requires knowing how the data would look in the absence of a kinship effect. For a typical node permutation test, the null hypothesis is that there is no statistical relationship between kinship and association in the observed network. One problem, however, is that the observed network is the outcome of many confounded biological and methodological factors, besides the hypothesized effect of interest, that operate simultaneously to shape the observed network (Figure 1). Some of these processes, which we call “nuisance effects”, can cause elevated false positives (“statistically significant” effects that are not real). For instance, a researcher interested in social preferences might incorrectly infer a social preference that is actually driven by a shared site preference. An individual may seem less socially associated with groupmates than others only because it is harder to observe, harder to identify, uses a smaller subset of sampled locations, has its main home range outside the study area, or left the study population early. If individuals or dyads are not sampled equally (or do not have the opportunity to be), then the resulting network edges will represent not only the outcome of animal behaviour, but also the differences in how individuals have been recorded. 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 theoretically provide equal sampling across individuals are not free of sampling biases if animal-borne proximity sensors vary in their sensitivity (e.g. due to tiny differences in soldering) or if some barcodes are more difficult to identify by computer vision. Methodological biases will vary in magnitude and importance across study designs, but they are arguably inevitable even under controlled conditions. Importantly, they can be difficult to control for using a node permutation test, making node permutations prone to elevated statistical error rates of this kind.

Pre-network permutations (or datastream permutations) can control for multiple potential nuisance effects, but they are prone to a different kind of elevated error rate. By constraining permutations to occur within blocks of time or space, pre-network permutations generate

randomized expected datasets that hold constant many features of the observed data including group sizes, number of observations per individual, 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 across space and time, and differences in sampling effort across space and time. However, in comparison with node permutations, pre-network permutations test a different null hypothesis: that the network structure is random after removing the effect of kinship and the specified nuisance effects. The networks expected under this null hypothesis often fail to maintain socially realistic network-level properties of real animal societies, such as the variance in edge weight or distribution of degree values in the permuted networks (Puga-Gonzalez, Sueur & Sosa 2020; Weiss et al. 2021). Thus, pre-network permutation tests potentially produce elevated false positives when a social network has other sources of social structure that are unrelated to the hypothesis of interest (Figure 1).

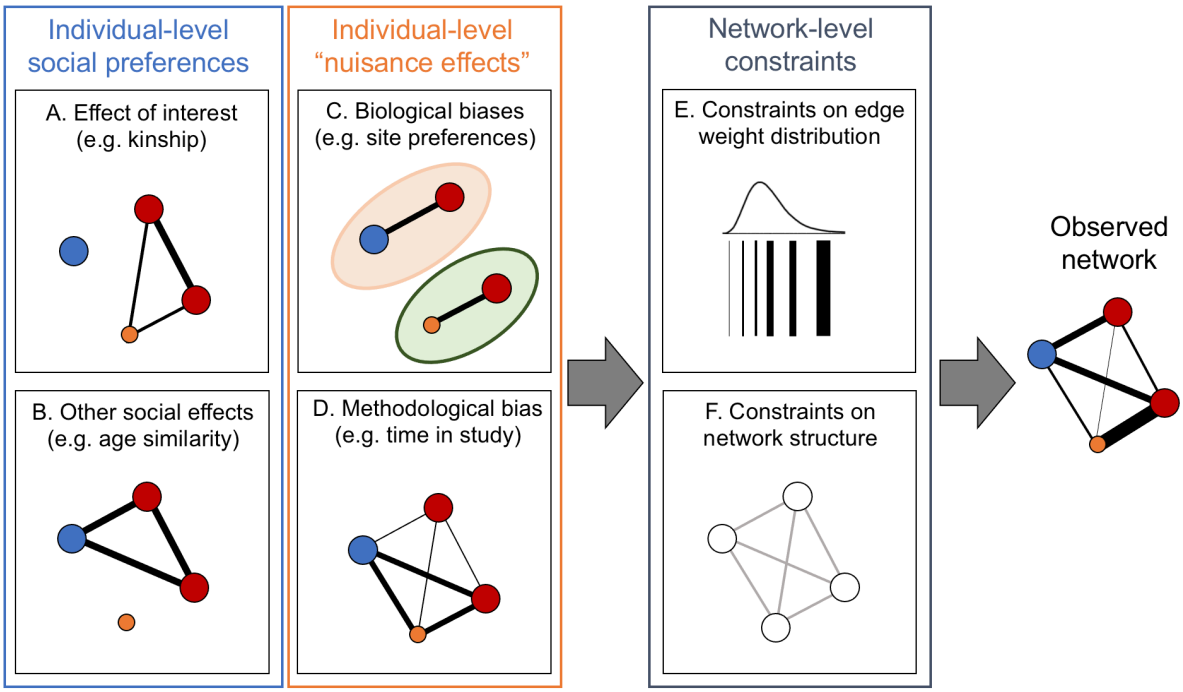


Figure 1. Observed social networks of association are shaped by factors besides the effect of interest. In this example showing only four nodes for simplicity, individuals prefer to associate with kin (node color in A) and with individuals of similar age (node size in B). Individuals also prefer certain sites (shaded ovals in C), and some individuals were observed less often (top node in D). These are illustrative examples; many other nuisance effects are possible in reality. All these effects collectively constrain network-level properties, such as the distribution of edge weights (E) or node degrees (F). Detecting the social preference for kin (A) therefore requires accounting for the consequences of other social preferences (B) and various nuisance effects (C, D) on the observed network. Pre-network (datastream) permutations create expected networks that preserve (control for) nuisance effects (C, D), but they also remove all social preferences, and thus fail to preserve socially realistic network-level traits (E, F), risking a Type I error. Node permutations create expected networks that preserve the network-level traits of the observed network; however, they fail to control for nuisance effects at the individual-level (C, D), so the observed social network might not accurately reflect the factors of interest, elevating Type I or II errors.

Biological and methodological nuisance effects can have interacting effects (see *Illustrating the Drivers of Type I and II Error* in Appendix). Even a biological process that is relevant to the hypothesis of interest (e.g. animal decisions about where to settle in space) can contribute to inaccuracies in the observed network. For example, compared to individuals at the centre of a study area, individuals at

the edge might have many associations with individuals that were never observable. Because node permutations assume the observed network is representative of actual association rates, 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 2021), and unfortunately, many of these biases can be quite difficult to correct using traditional correction terms in a statistical model (such as random effects or covariates). Capturing all nuisance effects in the model becomes increasingly challenging as the number of interacting effects increases, or when nuisance effects do not have regular patterns. For example, fitting individuals' location as a random effect requires assigning individuals to a singular spatial location, which is difficult when home ranges are continuously distributed and overlapping in space. The roles of biological and methodological nuisance effects on the observed network structure can be difficult to identify and disentangle, and failing to do so can easily lead to spurious outcomes (Farine & Aplin 2019). A major challenge, as it stands, is developing permutation methods that can test a hypothesis while robustly accounting for both social structure and nuisance effects. Our goal here is to propose an initial solution that can maintain rates of both false positives and false negatives to around 5% under a range of scenarios.

THE DOUBLE PERMUTATION METHOD

Our double permutation test (Figure 2) uses both pre-network and node permutations for what they each do best: pre-network permutations to account for nuisance effects by constraining swaps of observations within blocks of time and space (Whitehead, Bejder & Ottensmeyer 2005; Whitehead 2008; Sundaresan, Fischhoff & Dushoff 2009; Spiegel *et al.* 2016; Farine 2017), and node permutations to test for the effects of X on Y while holding social structure constant. We use the pre-network permutation to estimate the deviation of each unit's observed measure from its expected random value (e.g. the difference between a node's observed degree and the median values of the same node's degree across the permuted networks). We call this observed – expected difference the deviation score (e.g. Δdegree). These deviation scores, which account for nuisance effects, are then fit into a model of interest to generate a test statistic, and a node permutation test is then used to calculate the significance of the test statistic.

The double permutation approach is conceptually similar to generalised affiliation indices (Whitehead & James 2015), but to estimate the deviance of each unit's metric from random, it uses pre-network permutations rather than fitting covariates into regression model. Similar to extracting residuals from a model, our approach subtracts a measure of central tendency of the randomised data from the observed data because no single permuted network can be considered as the true expected random network. As a measure of central tendency for the expected values from pre-network permutations, we propose using the median because the distributions of expected values are often highly non-normal and long-tailed. However, the median assumes that the expected values have a unimodal distribution. To check this assumption, we recommend visualizing the distribution of expected values generated by the pre-network permutation (see also *Potential Improvements* in the Appendix).

The double permutation approach tests the null hypothesis that there is no relationship between a predictor and deviations from random social structure (within temporal and spatial constraints). The procedure can be applied to any model for calculating test statistics, such as those generated from regressions, Mantel tests (Mantel 1967), network regression models like MRAQP (Dekker, Krackhardt & Snijders 2007), and metrics such as the assortativity coefficient (Farine 2014).

We show that it also performs equally well with 'gambit-of-the-group' association data and with data collected using focal observations. We acknowledge that this is only one potential solution, and we therefore also highlight alternative methods that are also worth evaluating further, including several that are not based on permutation tests.

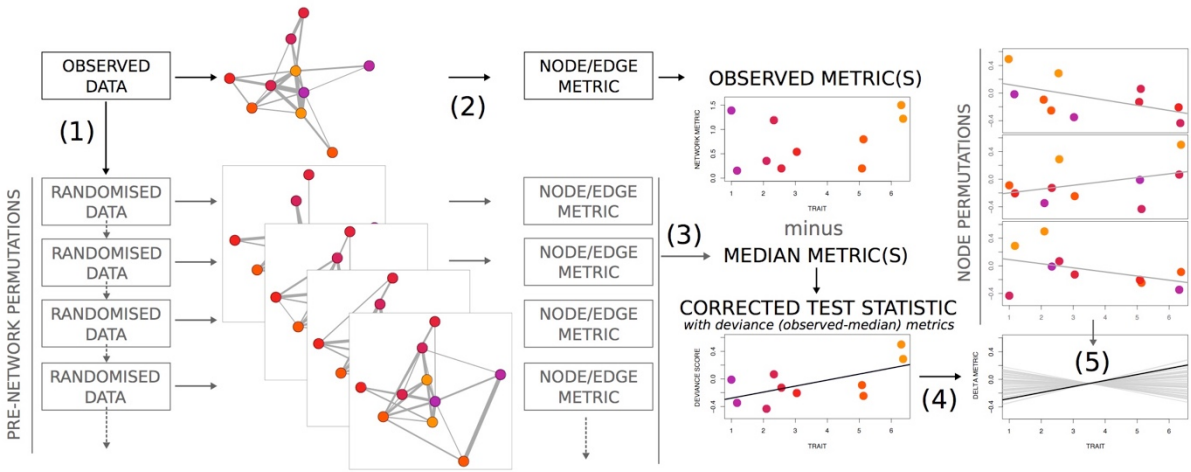


Figure 2. Overview of the double permutation method. Pre-network permutations (1) generate a null distribution of expected metric values for a unit of interest (e.g. a node's degree or an edge's weight) alongside each unit's observed metric (2). To remove nuisance effects for each unit, the median of expected metric values is subtracted from the observed metric value, yielding a corrected metric value called the "deviation score" (3). To generate a corrected test statistic, a test of interest (e.g. a linear model) is conducted with the deviation scores (4). To generate a P value, a node permutation test is then used to compare the corrected and expected test statistics (5).

TESTING THE ROBUSTNESS OF THE DOUBLE PERMUTATION APPROACH

We evaluate the performance of the double permutation tests when using gambit-of-the-group data (simulation 1), focal sampling data (simulation 2), and dyadic observations (simulation 3), both in the absence of any real relationship and when there is a strong nuisance effects. In simulations 1 and 2, we simulate the process of testing for a relationship between an individual trait and a social network metric, using three common weighted metrics (see Farine & Whitehead 2015): weighted degree (or strength), eigenvector centrality, and betweenness. In simulation 3, we simulate the process of testing for a link between pairwise kinship and association rate in a species that exhibits other non-kinship-based social preferences. For each simulated dataset, we calculate P values using several tests: node permutations, node permutations in which the model controls for covariates (number of observations and, where possible, location), pre-network permutations, pre-network permutations on the t-statistic (the equivalent of scaling the predictions, as proposed by Weiss *et al.* 2021), and the double permutation method. We create the networks and conduct the permutation tests using the R package *asnipe* (Farine 2013) and use the package *sna* (Butts 2008) to calculate network metrics.

Simulation 1: Regression between a node metric and a trait using gambit of the group data

We first simulate a researcher using group-based observations to test whether individuals with a given trait value are more gregarious under three biological scenarios: (i) individuals choose groups at random, (ii) individual choice is predicted by a social trait (T_S), whereby individuals with higher trait values choose to be in larger groups, and (iii) individual choice is predicted by a spatial

preference (L_S), whereby individuals with higher trait values are not more social but prefer sites with more resources that also hold more individuals. The last scenario represents an alternative driver of gregariousness that is challenging to control in model-fitting or using node permutations by location (because individual spatial preferences cannot be reduced down to a single value). Such variation in spatial distribution of individuals is common, for example the number of great tits (*Parus major*) in Wytham Woods varied ten-fold across different parts of the woodland (Farine *et al.* 2015). In each scenario, we simulated 100 replications for varying combinations of network size (5 to 120 individuals) and mean numbers of observations per individual (5 to 40). For each simulation (one network), we extract β coefficients (slopes) and t statistics by fitting the model, weighted metric ~ trait (+ covariates where applicable), using the lm function in R. Note that pre-network permutations inherently control for the number of observations of each individual, and we also control for where observations were made by restricting swaps to within locations. Further details for simulation 1 are in the Appendix.

Simulation 2: Estimating differences in group means using focal sampling data

We next simulate a researcher using focal sampling data to test whether different classes of individuals vary in their gregariousness. We build on the code of Puga-Gonzalez, Sueur and Sosa (2021) and Farine (2017), whereby each simulation assigns individuals to observations, with each observation having a focal individual and connections made between the focal and each individual observed. Simulations can be run with and without a sex difference in gregariousness. When the difference is present, females are made more gregarious by being disproportionately present in observations containing many individuals. 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 on the right wing and are only identifiable if their right side is observed, 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 was designed to allow an estimate of false positive rates (when no effect should be present but one is detected) and false negative rates (when an effect is present, but masked by the observation bias, and therefore not detected), and of the effect sizes before and after the observation bias occurs. For each simulation (one network), we extract β coefficients and t statistics by fitting the model, weighted metric ~ sex (+ covariates), in the lm function in R.

Simulation 3: Regression between edge weight and a dyadic relationship such as kinship using dyadic observations

Finally, we simulate a researcher testing for a link between kinship and social network structure, where edges are association rates that are driven by both kinship and other social preferences. To evaluate the impact of other social preferences 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. Our simulations comprise two scenarios. In the first, association rates (edge weights) are independent of kinship, which we achieve by randomizing the kinship matrix after generating it. In the second, associations are kin-biased such that strongly-

bonded associates have higher kinship on average and weak associates and non-associates (missing edges) have on average lower kinship. For each type of social network and scenario, we created 'real' social networks and then create 'observed' social networks by simulating observations. We simulate 100 replications for combinations of network sizes, with 5 to 120 individuals ranging and the mean numbers of observations per individual ranging from 5 to 40. For each simulation (one network), we conduct a multiple regression quadratic assignment procedure (MRQAP) with the model: edge weight ~ kinship (+ covariates). Further details for simulation 3 are in the Appendix.

THE DOUBLE PERMUTATION APPROACH IS ROBUST TO TYPE I AND TYPE II ERRORS

Our simulations confirm that the double permutation method is robust to type I errors across a range of scenarios. In simulation 1 (Table 1), when a trait is linked to neither social nor spatial outcomes, pre-network permutation tests are prone to elevated false positives (type I error rate of 26%), consistent with previous studies (Evans, Fisher & Silk 2020; Puga-Gonzalez, Sueur & Sosa 2021; Weiss *et al.* 2021). We further show that the tendency for pre-network permutation tests to generate type I errors is greater in smaller networks and when more data are collected (Figure S2, see also Evans, Fisher & Silk 2020). Pre-network permutations using the t value as a test statistic (which produces results equivalent to scaling the predictors) are also prone to false positives, and perform relatively poorly at detecting a real effect (detecting true effects consistently less often than other approaches). Node permutation tests perform particularly poorly when the effects are driven by non-social nuisance effects, such as variation in the spatial distribution of individuals, even when controlling for individuals' locations in the model. The double permutation test performs generally better than other methods across the three scenarios, producing conservative P values when no effect is present, detecting the relationship between a trait and social metrics when present, and being more conservative than other tests when the effect is driven by non-social factors. All tests appeared to struggle with confounded measures of eigenvector centrality and betweenness, but this is likely driven by the simulations not always producing datasets in which eigenvector centrality was strongly linked to spatial location, thereby largely over-estimating the false positive rate. However, because all methods were tested on identical simulated datasets, their relative performance can be meaningfully compared.

		No effects (should be low)	**Social effect only (should be high)**	**Spatial confound (should be low)**
Degree	**Node permutation (β)**	4.8%	88.1%	88.2%
	Controlled node permutation (β)	5.0%	89.9%	47.8%
	Pre-network permutation (β)	26.5%	90.5%	23.1%
	Pre-network permutation (t)	13.5%	65.1%	29.3%
	Double permutation	4.8%	89.0%	10.1%
Eigenvector Centrality	**Node permutation (β)**	4.7%	91.7%	91.9%
	Controlled node permutation (β)	4.6%	92.9%	60.6%

	Pre-network permutation (β)	33.9%	95.4%	36.6%
	Pre-network permutation (t)	14.5%	79.7%	21.5%
	Double permutation	4.8%	92.9%	29.2%
Betweenness	**Node permutation (β)**	5.1%	85.0%	85.5%
	Controlled node permutation (β)	5.2%	83.9%	73.6%
	Pre-network permutation (β)	20.1%	77.8%	72.1%
	Pre-network permutation (t)	10.5%	74.8%	72.2%
	Double permutation	5.0%	74.8%	69.6%

Table 1. Propensity for permutation tests to yield errors or detect real effects when using regression models to test hypotheses on networks collected using gambit-of-the-group 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 no effects are present, the expected proportion of significant results should be 5%. When a social effect is present, most results should be significant. When a spatial confound is present, the proportion of significant results should again approach 5%. Controlled node permutations include the number of observations in the models plus individuals' most common location in the spatial confound condition. Supplementary Figures S2-S4 show how the proportion of significant results is affected by the number of observations and the number of nodes in the network.

In simulation 2 (Table 2), pre-network permutations again show elevated type I error rates in the absence of true difference between classes of individuals and node permutations again show elevated type I error rates in the presence of nuisance effects. Controlling for nuisance effects in the model with node permutations helps under some circumstances (when there is a strong effect and an observation bias), but not others (e.g. when there is no effect and a bias). The double permutation test almost always performs as expected. One exception is the high levels of type II errors for betweenness. This suggests that because betweenness is an unstable metric (i.e. adding one edge can substantially change the distribution of betweenness values in the whole network) and pre-network permutations are perhaps not generating a meaningful null distribution on which the node permutation can operate.

		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)**
Degree	**Node permutation (β)**	4.2%	2.2%	60.0%	38.2%
	Controlled node permutation (β)	4.0%	19.8%	68.6%	4.6%
	Pre-network permutation (β)	32.2%	5.6%	34.4%	6.4%
	Pre-network permutation (t)	29.8%	69.8%	55.4%	34.6%

	Double permutation	4.8%	9.0%	5.6%	11.8%
Eigenvector centrality	**Node permutation (β)**	3.2%	1.6%	55.4%	33.0%
	Controlled node permutation (β)	4.8%	24.6%	69.0%	4.4%
	Pre-network permutation (β)	43.0%	1.4%	43.2%	1.8%
	Pre-network permutation (t)	20.4%	78.6%	62.4%	31.6%
	Double permutation	3.8%	3.0%	4.8%	7.2%
Betweenness	**Node permutation (β)**	6.4%	9.6%	67.6%	52.0%
	Controlled node permutation (β)	4.0%	36.0%	35.3%	25.4%
	Pre-network permutation (β)	14.8%	66.0%	16.6%	76.2%
	Pre-network permutation (t)	13.5%	68.8%	14.2%	79.7%
	Double permutation	5.8%	84.6%	7.6%	92.2%

Table 2. Propensity for permutation tests to produce type I and type II errors from datasets simulating focal sampling (model 2). Simulations comprise 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).

Simulation 3 (Table 3) shows that the double permutation test can reliably test null hypotheses that assume nonrandom social structure (similar to a node permutation test), such as whether association rates are predicted by kinship in cases where kinship effects are present and absent. As with the above two simulations, the double permutation test performs well when no real effect is present (i.e. type I error rates were close to 5%). All models in this simulation have elevated type II error rates because not all simulated networks with the added effect actually resulted in data with a clear effect, but the double permutation test performs more conservatively than node permutations (producing more type II errors).

	Kinship ≠ Associations (Type I errors)	**Kinship ∝ Associations (Type II errors)**
Node permutation (β)	5.1%	12.5%
Controlled node permutation (β)	4.9%	12.0%
Pre-network permutation (β)	18.6%	6.2%
Pre-network permutation (t)	3.0%	79.1%

Double permutation	5.2%	16.5%

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 S5-S6 show how the proportion of significant results is affected by the number of observations and the number of nodes in the network. 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 S6).

RECOMMENDATIONS

Our results suggest that across scenarios, the double permutation test often performs better than other approaches, particularly when there are strong confounds, such as sampling biases or spatial preference drivers, which create high rates of false positives in tests that do not include pre-network permutations. However, the double permutation test can be prone to being too conservative when applied to unstable global node metrics such as betweenness (Table 2), which might be better studied using node permutations (in the absence of confounding effects). Because the suitability of permutation methods has not been exhaustively tested with measures of betweenness, we cannot recommend any solution for hypothesis testing with betweenness in the presence of nuisance effects. Our recommendations across node metrics and scenarios are summarised in Table 4.

	Metric			
	Degree Centrality	**Eigenvector Centrality**	**Betweenness Centrality**	**Relational (edge weight)**
No clear nuisance effects	node or double	node or double	node	node or double
Nuisance effects expected	double	double	*unclear*	double

Table 4. Recommendations for type of permutations from simulations. In the absence of nuisance effects, or when weak nuisance effects can effectively be controlled in a node permutation (e.g. using restricted node permutations), then either node permutations or double permutations are likely to provide robust inference for most local network metrics (e.g. degree, eigenvector centrality) or for relational data (tests on edge weights). In the presence of nuisance effects, double permutations tests are generally recommended, except for betweenness.

THE CHALLENGE OF CALCULATING EFFECT SIZES

Inference will always benefit from relying less on *P* values and instead focusing more on estimating the magnitude and confidence of effect sizes (Nakagawa & Cuthill 2007), and the same is true for network data (Webber, Schneider & Vander Wal 2020). Franks *et al.* (2021) show that, under certain conditions, 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, not just their significance. We explore this using simulation 2, where one category of individuals is both more social and less observable. Such a situation is quite common in nature, for example female birds that are more drab than males (e.g. most sexually

dimorphic birds), shyer birds that are more social and more difficult to detect (Aplin *et al.* 2013), subadults that are more central but more difficult to recognise than adults (e.g. in primates where they have fewer markings or in the guineafowl example cited above), or if individuals vary in their habitat preferences [e.g. striated thornbills *Acanthiza lineata* are much more arborial than other birds they flock with, such as buff-rumped thornbills *Acanthiza reguloides*, and also much more camouflaged than flock-mates such as scarlet robins *Petroica boodang*, resulting in a lower average number of observations per individual (Farine & Milburn 2013)]. Simulation 2 has been used in several studies to estimate performance of permutation approaches (Farine 2017; Puga-Gonzalez, Sueur & Sosa 2021) as it can capture both the pre- and post-biased metrics.

While the original coefficient (before the observation bias) and estimated coefficient (with the observation bias) are correlated (r=0.54), controlling for the number of observations of each individual consistently inflates the estimated coefficient size (Figure S7). We test whether regression models can recover the original coefficient value using two approaches to fitting the number of observations as a covariate. First, we use a naïve model, where the scaled number of observations is simply added as a covariate. Second, we use a more informed model where 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). As expected, the naïve model performs worse, producing estimated effect sizes that are on average 1.8 times the original value (and up to 5.1 times the original value). Correctly fitting observations as an interaction term does 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 also perform very poorly at estimating effect sizes when the true effect is not present (the estimated effect sizes were on average over 250 times the true values, Figure S8).

One reason why regression models do not generate robust effect sizes when controlling for nuisance effects might be because they do not deal well with common situations problems with network data. One such situation is when individuals are observed in groups rather than in pairs, which is very common in social network studies (Sah, Mendez & Bansal 2019), and known to cause problems for analysis (Evans, Fisher & Silk 2020). For such gambit-of-the-group observations, the loss of each observation can result in a variable number of edges being removed within groups. For example, when using the simple association index to estimate proportion of time spent together (Hoppitt & Farine 2018), an individual A in a group of 10 that is imperfectly observed will have a reduction in degree of 0.9 for every 10% of decrease in its detection rate, while other perfectly detected members of the group will all decrease in degree by only 0.1 per 10% decrease in the detection rate of individual A. This difference in the effects of imperfect detections on degree among members of the same group become greater as group size increase, and is not resolved by increasing sampling. 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 further approaches.

Restricted node permutations

For many studies, it may 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. Say, for example, that 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. In this case, permutations could be restricted to only allow swaps among individuals that entered the study at approximately the same time. However, if multiple nuisance effects have to be accounted for, one can rapidly run out of sets of exchangeable individuals. For example, a study with 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 both parameters are binary, only 6-7 per class if one is trinary, and only 4-5 per class if both are trinary.

Study-specific simulations

How can we be sure that a chosen method is effective? One approach is to explicitly test how sensitive a given dataset is to generating false positives or false negatives under different hypothesis-testing approaches. 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. 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, one can observe the proportion of the tests that incorrectly produce a significant P value. This result can be also reported as evidence of the selected method's reliability. It is worth exploring further 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 and running a pre-network permutation test (and repeating these two steps many times) might provide an even more precise estimation of the true false positive rate, as it will be fully conditioned on the real observation data.

Multiple null models

Rather than testing one null hypothesis, which encourages confirmation bias, strong inference (Platt 1964) requires considering and testing multiple alternative hypotheses. Multiple null models can be used to collectively examine the different processes that might be shaping the patterns present in observation data, and can be highly informative. For example, 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). Multiple null models are particularly effective for generating an understanding of the effects of social decisions versus space use on social network structure (e.g. Figure S1). While it is important to control for the contribution of 'nuisance' spatial effects to social network structure when testing hypotheses about social decision-making, the process by which animals use space (and its links to social structure) is itself an important biological question (Webber & Vander Wal 2018; He, Maldonado-Chaparro & Farine 2019). We show an example of this in Figure S1, where both social and spatial processes shape the differences in the social connectedness of males and females. 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. Implementing multiple null models requires careful consideration of elevation in false positives (Webber, Schneider & Vander Wal 2020).

Bootstrapping (and its limitations)

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 2021), which 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 many cases these zero values could actually be weak associations that exist in the real world but were not observed. In such cases, bootstrapping the edge weights suggests that these unobserved edges have no uncertainty, which is obviously false. Thus, bootstrapping social network data should only be used with care.

Other non-permutation approaches

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).

Replicated networks

Pre-network permutation tests were initially designed to evaluate whether the social structure of a population is nonrandom, given sparse association data (Bejder, Fletcher & Brager 1998). Error rates for regression-based hypotheses can predictably degrade as the size of the observational dataset increases (Figure S2). The good news is that large datasets allow the creation of replicated networks (Hobson, Avery & Wright 2013), which is achieved by splitting the dataset to produce multiple networks (without overlapping observations). When doing so, it is important that each replicated 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.* 2021), each day (e.g. Boogert, Farine & Spencer 2014), or even each second (e.g. Blonder & Dornhaus 2011).

Networks that produce consistent results when tested independently provides much stronger support for a given hypothesis than any single network can. If, instead, effects are unstable over time, this might suggest either the presence of other underlying dynamics that warrant further investigation, more careful analyses, or the need for longer time periods for each replication.

Any inference becomes stronger again if each of the replicate social networks contains different sets of individuals. This could be if 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). Alternatively, the hypotheses could be tested on alternative subsets of the populations, such as different communities (Bond *et al.* 2020; Bond *et al.* 2021). Given sufficient data, replicated 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, as 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 hypothesised effects were in the same direction (binomial $P<0.0001$).

CONCLUSIONS

In this paper, we propose an approach that can avoid elevated false positives when either (1) using pre-network permutations to test hypotheses with networks that have structure unrelated to the hypothesis of interest or (2) using node permutations to test hypotheses in the presence of nuisance effects that generate structure in the data. 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 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 ignoring model assumptions or choosing an incorrect model structure (e.g. by failing to fit random slopes to a mixed effects model, see Schielzeth & Forstmeier 2009), and the same pitfalls occur when using these methods to calculate 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 count data (Ives 2015) or not (O'Hara & Kotze 2010). In the context of social network data, each permutation procedure (including constricting the swaps in different ways) tests a specific null hypothesis (see Figure S1 for an example), so the crucial statistical consideration is ensuring that the correct null hypothesis is being tested.

One particularly important point highlighted by our work, and that of others (e.g. Franks *et al.* 2021), is the need to pay particularly close attention to the importance of different processes for a given hypothesis of interest. For example, 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) could be considered a nuisance effect. However, if we aim to study the transmission of information or pathogens, then space use is an important factor of interest contributing to the outcome of the transmission process. Thus, a given process 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.* 2021).

In contrast to parametric approaches (or drawing heavily from these when producing a test statistic), permutation tests 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. Thus, one need not assume that the observed network corresponds exactly to the true network of interest (e.g. social structure). Using permutation tests instead requires and encourages a focus on thinking carefully about what specific processes may have produced the patterns in a given observed dataset. Evaluating the relative contribution of different processes can even be done by using a range of permutation tests to measure the relative deviations of the observed data from different null expectations. The strength and robustness of permutation tests therefore lies in their flexibility and simplicity.

ACKNOWLEDGEMENTS

We thank the Farine and Carter labs, Josh Firth, Matt Silk, Michael Weiss, Dan Franks, Lauren Brent, Sebastian Sosa, and the many researchers who contacted the authors with questions about the pre-print, for helpful comments, insights, and discussions regarding the issues presented in this paper. DRF was funded by the Max Planck Society, an Eccellenza Professorship Grant of the Swiss National Science Foundation (Grant Number PCEFP3_187058), 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 support from the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – EXC 2117 – 422037984.

AUTHOR CONTRIBUTIONS

DRF AND GGC developed the method, and wrote the manuscript together. DRF developed and ran the simulations.

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., Farine, D.R., Morand-Ferron, J., Cole, E.F., Cockburn, A. & Sheldon, B.C. (2013) Individual personalities predict social behaviour in wild networks of great tits (Parus major). *Ecology Letters,* **16,** 1365–1372.

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,** e20298.

Bond, M.L., Konig, B., Lee, D.E., Ozgul, A. & Farine, D.R. (2020) Proximity to humans affects local social structure in a giraffe metapopulation. *Journal of Animal Ecology*.

Bond, M.L., Lee, D.E., Farine, D.R., Ozgul, A. & Konig, B. (2021) Sociality increases survival in adult giraffes. *Proceedings of the Royal Society B,* **288,** 20202770.

Bonnell, T.R. & Vilette, C. (2021) Constructing and analysing time-aggregated networks: The role of bootstrapping, permutation and simulation. *Methods in Ecology and Evolution,* **10.1111/2041-210X.13351**.

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,** 20182579.

Butts, C.T. (2008) Social Network Analysis with sna. *Journal of Statistical Software,* **24,** 1-50.

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. (2021) The importance of individual-to-society feedbacks in animal ecology and evolution. *Journal of Animal Ecology,* **90,** 27-44.

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. (2021) Testosterone-mediated behavior shapes the emergent properties of social networks. *Journal of Animal Ecology,* **90,** 131-142.

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.

Dekker, D., Krackhardt, D. & Snijders, T. (2007) Sensitivity of MRQAP tests to collinearity and autocorrelation conditions. *Psychometrika,* **72,** 563-581.

Dormann, C.F., Fründ, J., Blüthgen, N. & Gruber, B. (2009) Indices, Graphs and Null Models: Analyzing Bipartite Ecological Networks. *The Open Ecology Journal,,* **2,** 7-24.

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

Farine, D.R. (2013) Animal Social Network Inference and Permutations for Ecologists in R using asnipe. *Methods in Ecology and Evolution,* **4,** 1187–1194.

Farine, D.R. (2014) Measuring phenotypic assortment in animal social networks: weighted associations are more robust than binary edges. *Animal Behaviour,* **89,** 141-153.

Farine, D.R. (2017) A guide to null models for animal social network analysis. *Methods in Ecology and Evolution,* **8,** 1309-1320.

Farine, D.R. (2018) When to choose dynamic vs. static social network analysis. *Journal of Animal Ecology,* **87,** 128-138.

Farine, D.R. & Aplin, L.M. (2019) Spurious inference when comparing networks. *Proceedings of the National Academy of Sciences of the United States of America,* **116,** 16674-16675.

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., Psorakis, I., Radersma, R., Verhelst, B., Voelkl, B. & Sheldon, B.C. (2015) The role of social and ecological processes in structuring animal populations: a case study from automated tracking of wild birds. *Royal Society Open Science,* **2,** 150057.

Farine, D.R. & Milburn, P.J. (2013) Social organisation of thornbill-dominated mixed-species flocks using social network analysis. *Behavioral Ecology and Sociobiology,* **67,** 321-330.

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

Farine, D.R. & Whitehead, H. (2015) Constructing, conducting, and interpreting animal social network analysis. *Journal of Animal Ecology,* **84,** 1144-1163.
Firth, J.A., Cole, E.F., Ioannou, C.C., Quinn, J.L., Aplin, L.M., Culina, A., McMahon, K. & Sheldon, B.C. (2018) Personality shapes pair bonding in a wild bird social system. *Nature Ecology & Evolution,* **2,** 1696-1699.
Forstmeier, W. & Schielzeth, H. (2011) Cryptic multiple hypotheses testing in linear models: overestimated effect sizes and the winner's curse. *Behavioral Ecology and Sociobiology,* **65,** 47-55.
Forstmeier, W., Wagenmakers, E.J. & Parker, T.H. (2017) Detecting and avoiding likely false-positive findings - a practical guide. *Biological Reviews,* **92,** 1941-1968.
Franks, D.W., Weiss, M.N., Silk, M.J., Perryman, R.J.Y. & Croft, D.P. (2021) Calculating effect sizes in animal social network analysis. *Methods in Ecology and Evolution,* **12,** 33-41.
Gimenez, O., Mansilla, L., Klaich, M.J., Coscarella, M.A., Pedraza, S.N. & Crespo, E.A. (2019) Inferring animal social networks with imperfect detection. *Ecological Modelling,* **401,** 69-74.
Gotelli, N.J. & Graves, G.R. (1996) *Null models in ecology*. Smithsonian Institution Press, Washington, DC, USA.
Harvey, P.H. (1987) On the use of null hypotheses in biogeography. *Neutral models in biology* (ed. M.H. Nitechi, Hoffman, A.), pp. 109–118. Oxford University Press, New York.
He, P., Maldonado-Chaparro, A. & Farine, D.R. (2019) The role of habitat configuration in shaping social structure: a gap in studies of animal social complexity. *Behavioral Ecology and Sociobiology,* **73,** 9.
Hobson, E.A., Avery, M.L. & Wright, T.F. (2013) An analytical framework for quantifying and testing patterns of temporal dynamics in social networks. *Animal Behaviour,* **85,** 83-96.
Hobson, E.A., Monster, D. & Dedeo, S. (2019) Strategic heuristics underlie animal dominance hierarchies and provide evidence of group-level social knowledge. *arXiv***,** 1810.07215v07213.
Hoppitt, W. & Farine, D.R. (2018) Association indices for quantifying social relationships: how to deal with missing observations of individuals or groups. *Animal Behaviour,* **136,** 227-238.
Ives, A.R. (2015) For testing the significance of regression coefficients, go ahead and log-transform count data. *Methods in Ecology and Evolution,* **6,** 828-835.
Koskinen, J.H. & Snijders, T.A.B. (2007) Bayesian inference for dynamic social network data. *Journal of Statistical Planning and Inference,* **137,** 3930-3938.
Langen, T.A. (1996) Social learning of a novel foraging skill by white-throated magpie-jays (Calocitta formosa, Corvidae): A field experiment. *Ethology,* **102,** 157-166.
Lusseau, D., Whitehead, H. & Gero, S. (2008) Incorporating uncertainty into the study of animal social networks. *Animal Behaviour,* **75,** 1809-1815.
Mantel, N. (1967) The detection of disease clustering and a generalized regression approach. *Cancer Research,* **27.2,** 209-220.
Miller, E.T., Farine, D.R. & Trisos, C.H. (2017) Phylogenetic community structure metrics and null models: a review with new methods and software. *Ecography,* **40,** 461-477.
Nakagawa, S. & Cuthill, I.C. (2007) Effect size, confidence interval and statistical significance: a practical guide for biologists. *Biological Reviews,* **82,** 591-605.
O'Hara, R.B. & Kotze, D.J. (2010) Do not log-transform count data. *Methods in Ecology and Evolution,* **1,** 118-122.
Papageorgiou, D., Christensen, C., Gall, G.E., Klarevas-Irby, J.A., Nyaguthii, B., Couzin, I.D. & Farine, D.R. (2019) The multilevel society of a small-brained bird. *Current Biology,* **29,** R1120-R1121.
Platt, J.R. (1964) Strong Inference. *Science,* **146,** 347-353.
Puga-Gonzalez, I., Sueur, C. & Sosa, S. (2021) Null models for animal social network analysis and data collected via focal sampling: Pre-network or node network permutation? *Methods in Ecology and Evolution,* **12,** 22-32.
Puth, M.T., Neuhauser, M. & Ruxton, G.D. (2015) On the variety of methods for calculating confidence intervals by bootstrapping. *Journal of Animal Ecology,* **84,** 892-897.
Ripperger, S.P., Carter, G.G., Duda, N., Koelpin, A., Cassens, B., Kapitza, R., Josic, D., Berrio-Martinez, J., Page, R.A. & Mayer, F. (2019) Vampire Bats that Cooperate in the Lab Maintain Their Social Networks in the Wild. *Current Biology,* **29,** 4139-4144.
Ripperger, S.P., Carter, G.G., Page, R.A., Duda, N., Koelpin, A., Weigel, R., Hartmann, M., Nowak, T., Thielecke, J., Schadhauser, M., Robert, J., Herbst, S., Meyer-Wegener, K., Wagemann, P., Schroder-Preikschat, W., Cassens, B., Kapitza, R., Dressler, F. & Mayer, F. (2020) Thinking small: Next-generation sensor networks close the size gap in vertebrate biologging. *Plos Biology,* **18,** e3000655.
Ryder, T.B., Horton, B.M., van den Tillaart, M., Morales, J.D. & Moore, I.T. (2012) Proximity data-loggers increase the quantity and quality of social network data. *Biology Letters,* **8,** 917-920.
Sah, P., Mendez, J.D. & Bansal, S. (2019) A multi-species repository of social networks. *Scientific Data,* **6**.

Schielzeth, H. & Forstmeier, W. (2009) Conclusions beyond support: overconfident estimates in mixed models. *Behavioral Ecology,* **20,** 416-420.

Spiegel, O., Leu, S.T., Sih, A. & Bull, C.M. (2016) Socially-interacting or indifferent neighbors? Randomization of movement paths to tease apart social preference and spatial constraints. *Methods in Ecology and Evolution,* **7,** 971-979.

Sundaresan, S.R., Fischhoff, I.R. & Dushoff, J. (2009) Avoiding spurious findings of nonrandom social structure in association data. *Animal Behaviour,* **77,** 1381-1385.

Thomson, C.E., Winney, I.S., Salles, O.C. & Pujol, B. (2018) A guide to using a multiple-matrix animal model to disentangle genetic and nongenetic causes of phenotypic variance. *Plos One,* **13,** e0197720.

Webber, Q.M.R., Schneider, D.C. & Vander Wal, E. (2020) Is less more? A commentary on the practice of 'metric hacking' in animal social network analysis. *Animal Behaviour,* **168,** 109-120.

Webber, Q.M.R. & Vander Wal, E. (2018) An evolutionary framework outlining the integration of individual social and spatial ecology. *Journal of Animal Ecology,* **87,** 113-127.

Weiss, M.N., Franks, D.W., Brent, L.J.N., Ellis, S., Silk, M.J. & Croft, D.P. (2021) Common datastream permutations of animal social network data are not appropriate for hypothesis testing using regression models. *Methods in Ecology and Evolution,* **10.1111/2041-210X.13508**.

Whitehead, H. (2008) *Analyzing animal societies*. University of Chicago Press, Chicago, USA.

Whitehead, H., Bejder, L. & Ottensmeyer, C.A. (2005) Testing association patterns: issues arising and extensions. *Animal Behaviour,* **69,** e1-e6.

Whitehead, H. & James, R. (2015) Generalized affiliation indices extract affiliations from social network data. *Methods in Ecology and Evolution,* **6,** 836-844.

APPENDIX

ILLUSTRATING THE DRIVERS OF TYPE I AND TYPE II ERRORS

Here we clarify the main problem by considering a simple, verbal, but concrete example of why errors can arise when using permutation tests. Imagine a study population where animals cluster for warmth each night. Clustering associations include 2-10 individuals and the group composition can change from night to night. The dataset contains a list of observed clusters, their location, time, and the individuals in those clusters that could be correctly identified. From these data, researchers generate a network describing the propensity for each dyad to be observed in the same cluster with the aim of finding out if kin are more likely to cluster: do individuals preferentially cluster with kin?

First, the researchers consider a node permutation approach (e.g. using a Mantel or MRQAP test) to correlate pairwise kinship and association. However, kin can associate more than expected by chance without kin-biased behaviors or even kin discrimination—for example if siblings are born at the same time, limited to similar home ranges, and then disperse at around the same time. Under such a scenario, a significant result from a node permutation test would correctly support the hypothesis that the network is kin-structured, but it would be spurious support for the hypothesis that kinship is a causal driver of social associations (false positive). Kinship-biased association could be undetected or underestimated (false negative) if younger animals are both more likely to associate with kin but also less likely to be individually marked and recorded. In the presence of sampling bias or other nuisance effects, hypotheses about the process generating an effect of kinship on association could be challenging to accurately assess using basic node permutation tests (but see Alternative Approaches section for more discussion on how node permutations can be restricted to help alleviate some of these effects).

The researchers might therefore turn to using a pre-network permutation test. To create expected outcomes for the relationship between kinship and edge weights, they repeatedly swap individual observations within sampled locations and time periods. Doing so allows them to eliminate the nonsocial nuisance effects described above, but it also assumes that no other social factors cause social network structure, because pre-network permutations assume a null hypothesis of no social effects at all. Imagine that in this species, individuals always 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 (association rates) across all its possible associates, because most association rates are zeros while a few are larger (i.e. closer to one). This variance in edge weight is much greater than expected from random associations (few zeros but also few larger values). When the researchers create randomised networks using pre-network permutations, many zero association rates are replaced with non-zero values as individuals that never co-occurred are swapped into groups together, and the individuals in the resulting random networks almost never spend 90% of their time with a few individuals (i.e. the previously strong edges become weaker). Even if association rates are not kin-biased and the close social bonds are randomly distributed with regards to kinship, the extremely nonrandom social structure of the observed network could easily lead to a false positive with regards to the effect of kinship (Weiss *et al.* 2021). If even a few strong bonds exist among kin by chance in the observed data, these strong kin bonds might not ever appear in the randomized networks. 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.*

2021). Thus, once again, the actual effect of kinship on association is challenging to accurately evaluate in the presence of a nuisance effect (in this case social preferences unrelated to kinship).

A different potential problem with pre-network permutation tests is that they can provide overly-confident estimates of minor deviation from random. 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 in any hypothesis test a sufficiently large number of observations can invariably produce P values well below 0.05 even when the effect size is not biologically important.

Pre-network permutations can also suffer from incorrect inference when networks contain only a few nodes. For instance, consider a (very simplistic) network containing only three nodes (male, male and female), and with only three observed dyadic associations, 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). While any inference drawn from three observations of three individuals is immediately apparent as unreliable, the same problem can be present, and much harder to notice, in more complex analyses.

The previous example illustrates problems caused by nuisance effects when the variables are traits of edges (kinship and association). The same principles apply when testing hypotheses relating the traits of individuals. For instance, a false impression that individuals of one sex are more gregarious (higher network 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. 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, and 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 S1).

Although it is useful to conceptually distinguish between effects that are social versus nonsocial or biological versus methodological, these can be difficult to disentangle in practice because biological biases can influence methodological biases and social behaviours can contribute to the apparently 'non-social' effects. Studies could benefit from making more explicit considerations of the social decisions that contribute to spatial and temporal biases. For example, individuals' spatial ranges could be determined by 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.

Another example is how social behaviour can influence observability. Consider a study testing if male birds have higher degree centrality. If females tend to be found at the periphery of groups, then an observer standing close to the centre 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 centre 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.

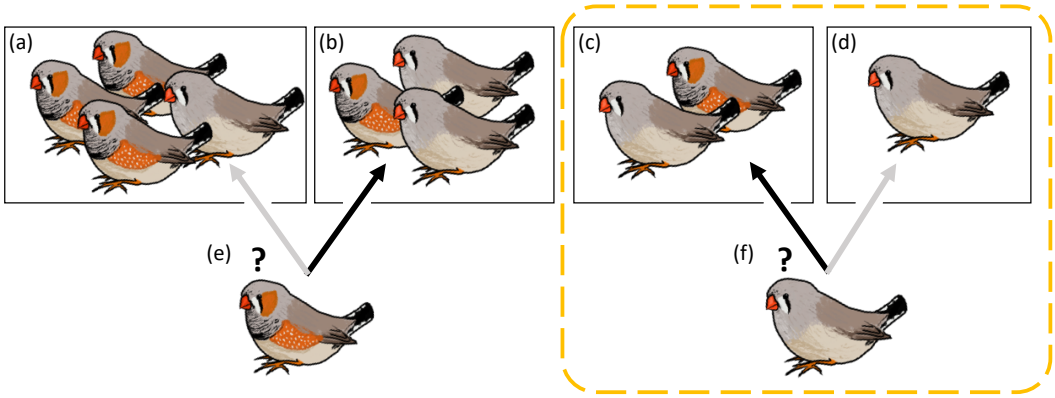


Figure S1. Example of the challenges associated with testing hypotheses on social data. Compared to females, are males more often in larger groups? In the scenario shown above, males (colourful individuals) are observed in larger groups (e.g. groups a, b, c) more often than are females (less colourful individuals). However, now consider a slightly different question: Compared to females, do males have a greater preference for larger groups? The same observation is not conclusive regarding the causal processes. Imagine that groups a and b are outside a territory (shown by the dashed line) defended by a single dominant male (shown in c) that allows females to enter freely and form groups (c and d) but excludes other males. Even if all males choose smaller groups (b vs a), and all females (e.g. f) choose larger groups (c vs d) given their available options, we could still observe males in larger groups than females due to the constraints imposed by the territory. Thus, even if males are equally or less gregarious, they can end up in larger groups and having more social connections. However, by permuting the individual observations within locations (a pre-network permutation), we can simulate individuals of both sexes making random social decisions 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. Because males would end up in larger groups in both the observed and permuted datasets, this pre-network permutation test would correctly indicate that males do not prefer larger groups. However, the same test would incorrectly fail to detect that males and females differ in their numbers of connections. The former question might be the focus of a study on how individual decisions shape network structure, whereas the latter question might be the focus of a study of how network structure shapes information or pathogen transmission.

An advantage of pre-network permutations is that they allow precise control over many possible nuisance effects, without needing to measure or even identify all of them. For example, if a lower proportion of individuals are individually-identifiable at the edge of a study area, then individuals seen at the edge would always occur in groups containing fewer individuals that can be individually-identified. In this case, controlling for space would automatically control for differences in group size that arise due to spatial variation in identification rates by only swapping individuals into other groups that also had fewer individual-identified individuals. Such an effect could be challenging to control for explicitly in a statistical model. On the other hand, pre-network permutation tests allow many other important aspects of the permuted networks to change (such as degree distributions or variances), and thus pre-network permutations alone should not be used to assess the effect of a predictor while controlling for social structure (Puga-Gonzalez, Sueur & Sosa 2021; Weiss *et al.* 2021). Ideally, one would use pre-network permutations to control for nuisance effects without making unrealistic random networks, which would include maintaining the variance in association strengths.

POTENTIAL IMPROVEMENTS OF THE DOUBLE PERMUTATION TEST

There are many further directions and refinements that can be explored for the double-permutation test. While we have demonstrated that the 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 2), 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 deviation score for each unit in the analysis (each node or each edge), and use these many measurements to estimate the 95% range of the corrected test statistic. Carrying uncertainty through the analysis could be implemented using a Bayesian framework (Farine & Strandburg-Peshkin 2015), or to model the dynamic process of the observation of connections in the network (Koskinen & Snijders 2007). Thus, there remains significant grounds for continued improvements in the methods for 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 nonrandom 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.

SIMULATION 1 DETAILS

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. Because the simulations did not always create a relationship (especially in simulations of eigenvector centrality and betweenness), we stochastically re-ordered the traits when $T_S = TRUE$ and no relationship was present to create a relationship. This was done by ordering individuals by their trait

value and probabilistically allocating them to nodes according to the node values, thereby resulting in a strong relationship.

From these observation data, we follow the design of our double permutation method (see Figure 2) to first calculate how each node's degree deviates from what is expected from random behaviour, and, second, to calculate the relationship between the deviance score for 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. In the first variant (location effect, $L_S = TRUE$), we assume that X corresponds to a spatial preference, 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, corresponding to a larger X, contain more resources and therefore can hold more individuals. The second variant (no location effect, $L_S = FALSE$) is a scenario in which the group size values X corresponds to the outcome of social decisions, such that individuals with a larger X have a preference for larger groups. 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 the decisions are social ($T_S = TRUE$ and $L_S = FALSE$) and in another scenario the relationship between T_i and D_i arises from decisions that are driven by a non-social covariate ($T_S = TRUE$ and $L_S = TRUE$). To control for the location X_j that each group was observed in when $L_S = TRUE$, we use within-location swaps in the pre-network permutation tests. We also calculate individual's most common location to include as a covariate in controlled node permutations.

SIMULATION 3 DETAILS

In simulation 3, we first create 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).

Second, we make an average of K observations per individual. Specifically, 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. To create associations 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 being observed together when they are both present to be 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. We select the number of individuals and observations (N and K) using the same parameter sets as in model 1.

Third, we 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 distributions assume that 1 corresponds to the closest relatives in the population. We use these parameters to create kinship distributions with differences in mean relative kinship (0.33, 0.5, 0.6, respectively) according to social relationship type. Note that these relative kinship values can be divided by two to create a maximum kinship value of 0.5 with no effect on the outputs of the model.

Our simulations comprise two scenarios. In the first scenario, edge weights are unrelated to kinship, which we achieve by randomizing the kinship matrix after generating it (thereby keeping the same relationship between the variance in the network edges and in the kinship matrix). In the

second scenario, associations are kin-biased by keeping the kinship matrix as it was generated, thus strongly-bonded associates have the highest kinship on average and weak associates and non-associates (missing edges) have on average the lowest kinship. We create the networks, conduct the pre-network permutation tests, and conduct the regressions using the MRQAP functionality in the R package *asnipe* (Farine 2013).

SUPPLEMENTARY FIGURES

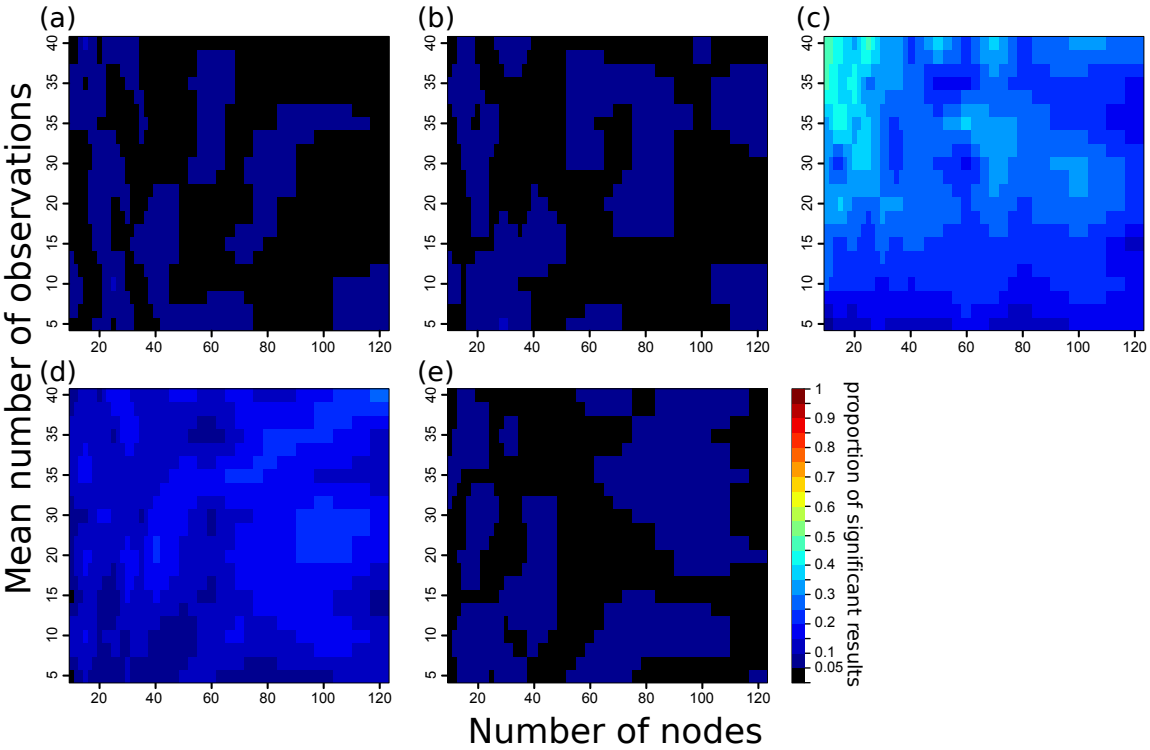


Figure S2. Proportions of statistically significant results for differently-sized networks and different observation efforts for simulation 1 under the scenario when $T_S = FALSE$ (no effect is present) and $L_S = FALSE$ (individuals are not preferentially located in different patches), with degree as the network metric. 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 but many observations (top left of the plot).

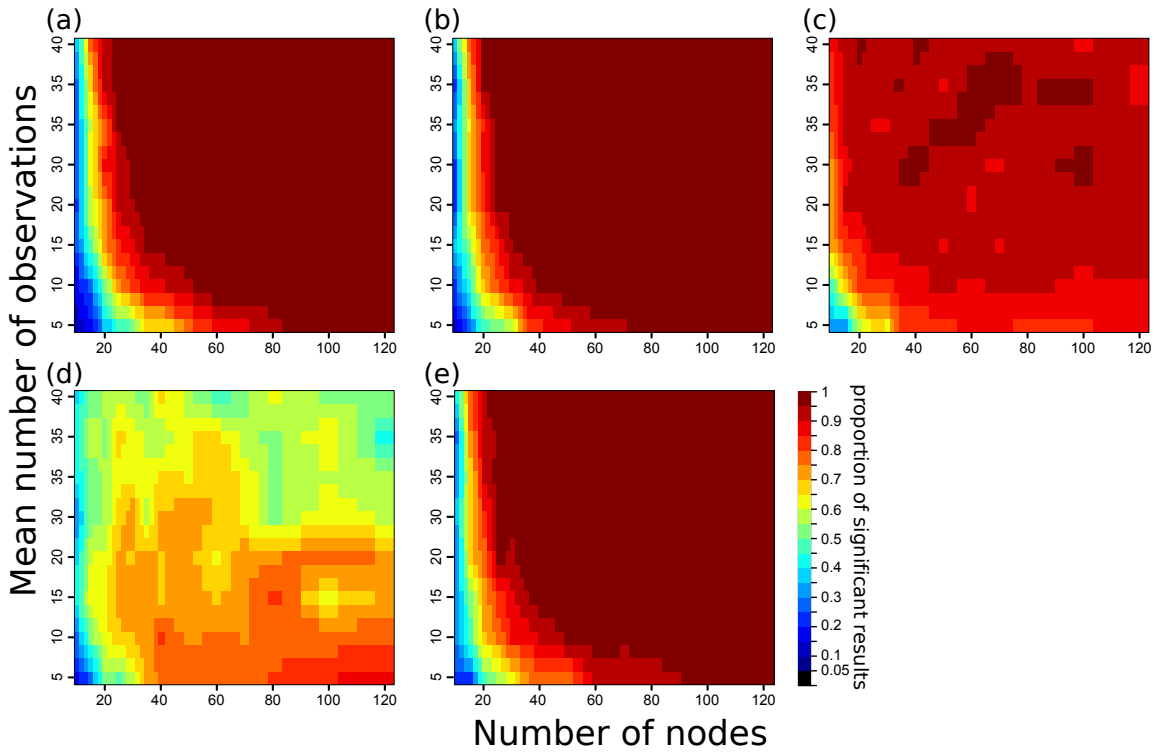


Figure S3. Proportions of statistically significant results for differently-sized networks and different observation efforts for simulation 1 under the scenario when $T_S = TRUE$ (an effect is present) and $L_S = FALSE$ (the effect is not a spatial confound), with degree as the network metric. Panels 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).

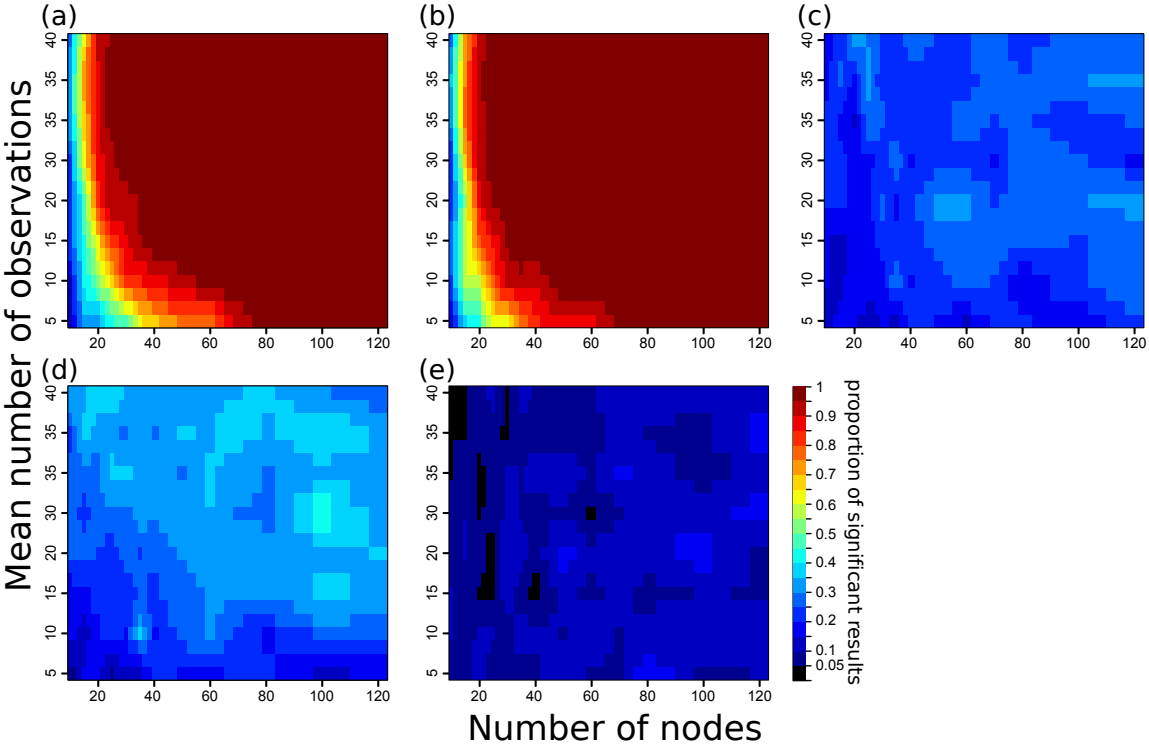


Figure S4. Proportions of statistically significant results for differently-sized networks and different observation efforts for simulation 1 under the scenario when $T_S = TRUE$ (an effect is present) and $L_S = TRUE$ (the effect is a spatial confound), with degree as the network metric. Panels represent 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).

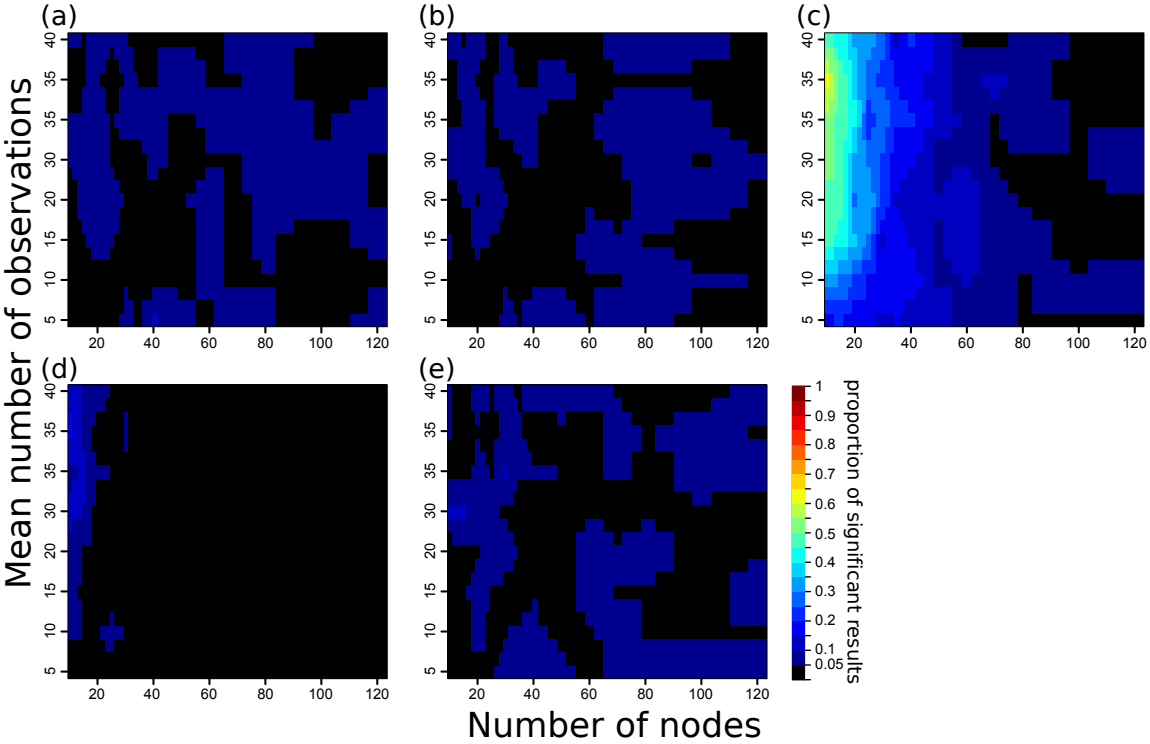


Figure S5. Proportions of statistically significant results for differently-sized networks and different observation efforts for simulation 3 under the scenario where kinship does not predict association rates (edge weights). Panels represent the P values calculated using (a) node permutation tests with 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).

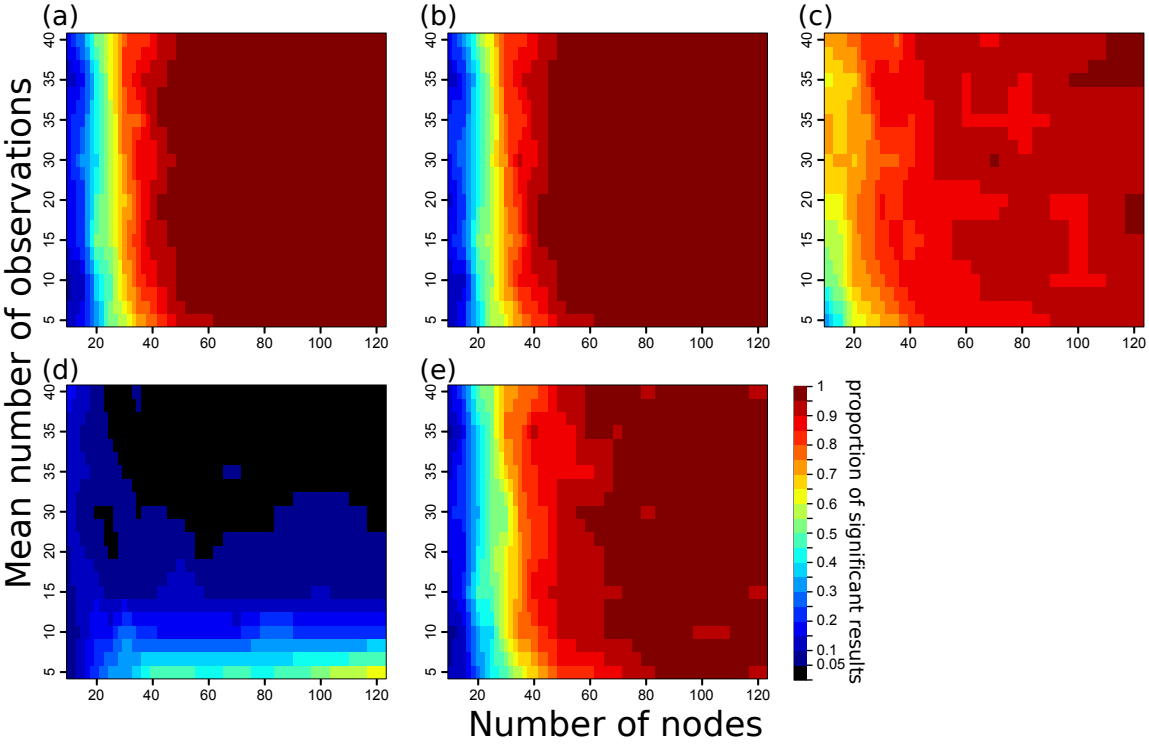


Figure S6. Proportions of statistically significant results for differently-sized networks and different observation efforts for simulation 3 under the scenario where kinship predicts association rates (edge weights). Panels represent the P values calculated using (a) node permutation tests with the 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).

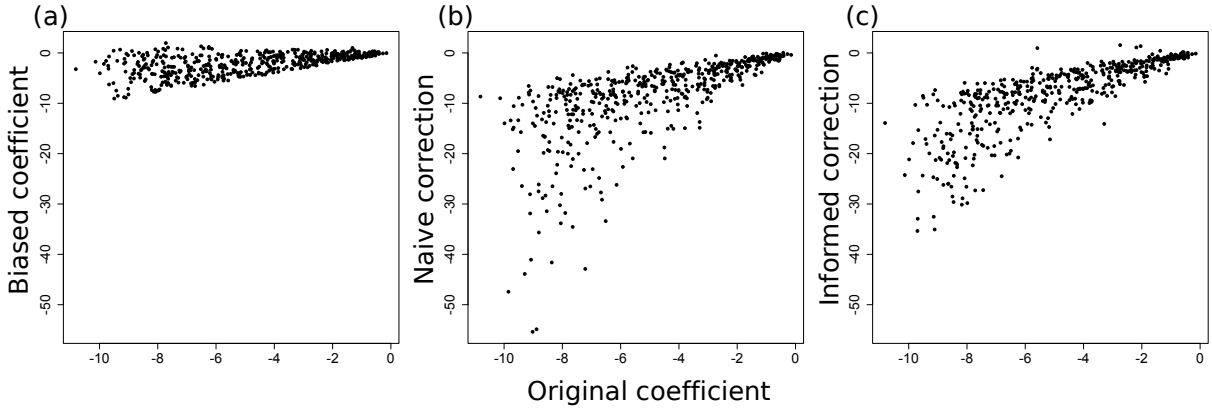


Figure S7. Relationship between the original coefficient value (the relationship between degree and sex prior to introducing an observation bias) and estimations of the coefficient value using the data from simulation 2 where an effect is present (females are more gregarious). (a) The original coefficient 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.

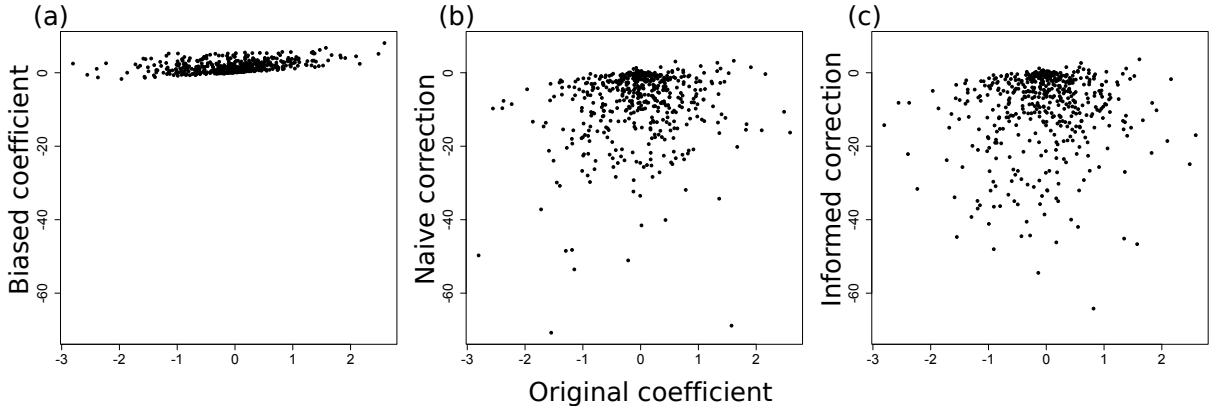
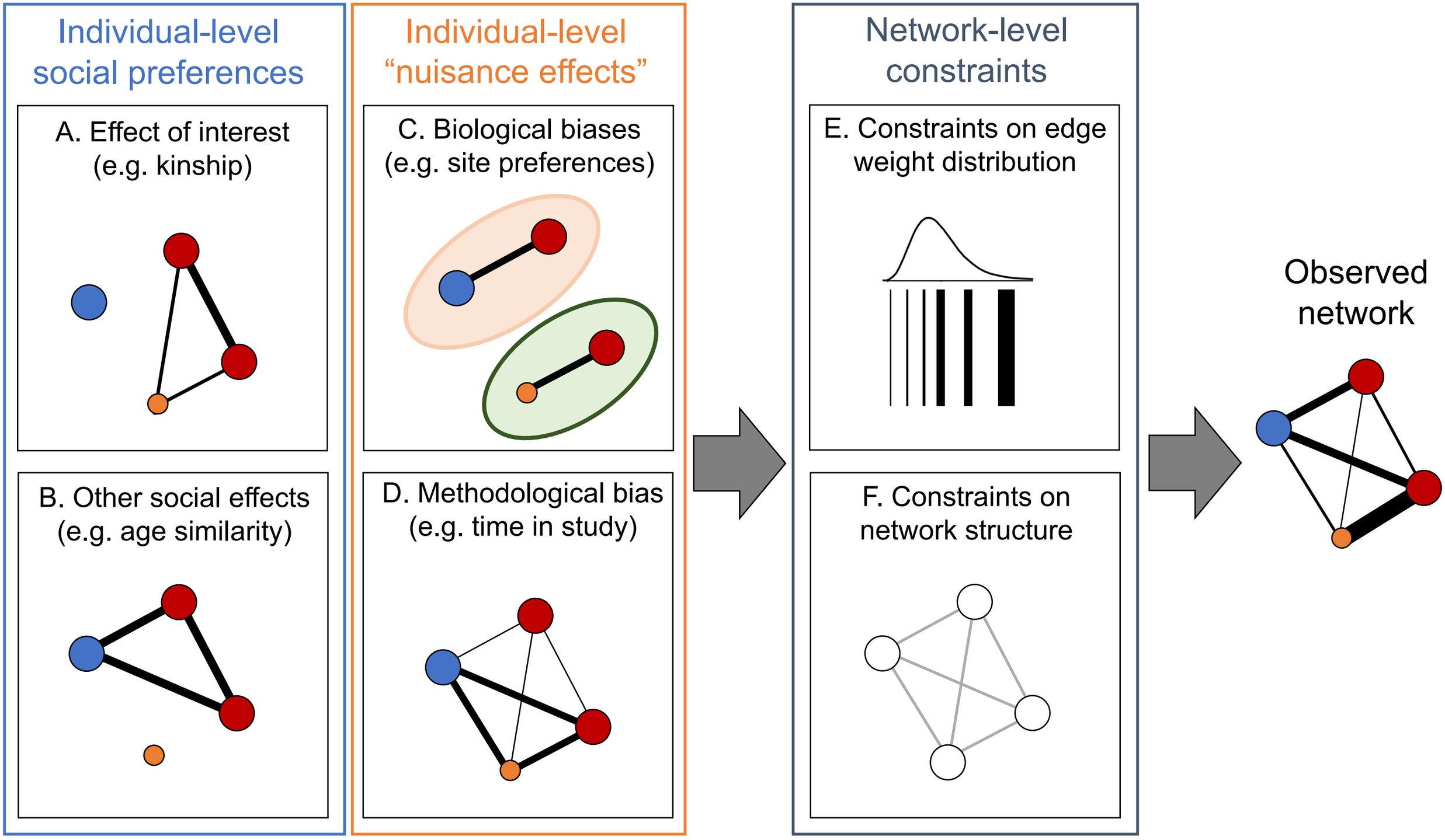
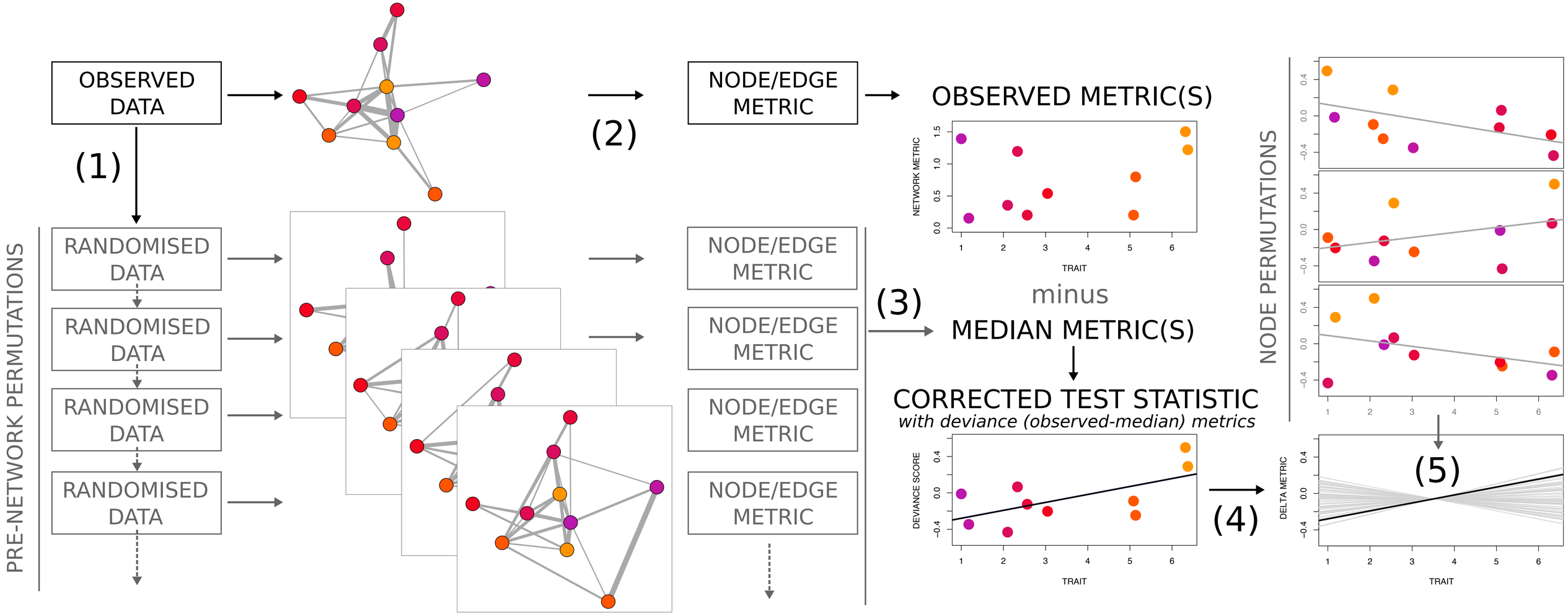


Figure S8. Relationship between the original coefficient value (the relationship between degree and sex prior to introducing an observation bias) and estimations of the coefficient value using the data from simulation 2 where no effect is present (females and males are equally gregarious). (a) The original 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.


Individual-level social preferences
A. Effect of interest (e.g. kinship)
B. Other social effects (e.g. age similarity)
Individual-level "nuisance effects"
C. Biological biases (e.g. site preferences)
D. Methodological bias (e.g. time in study)
Network-level constraints
E. Constraints on edge weight distribution
F. Constraints on network structure
Observed network


OBSERVED DATA
(1)
(2)
NODE/EDGE METRIC
OBSERVED METRIC(S)
NETWORK METRIC
TRAIT
PRE-NETWORK PERMUTATIONS
RANDOMISED DATA
RANDOMISED DATA
RANDOMISED DATA
RANDOMISED DATA
NODE/EDGE METRIC
NODE/EDGE METRIC
NODE/EDGE METRIC
NODE/EDGE METRIC
(3)
minus
MEDIAN METRIC(S)
CORRECTED TEST STATISTIC
with deviance (observed-median) metrics
DEVIANCE SCORE
TRAIT
(4)
NODE PERMUTATIONS
(5)
DELTA METRIC
TRAIT