Association indices for quantifying social relationships: how to deal with missing observations of individuals or groups

William Hoppitt\textsuperscript{1} and Damien Farine\textsuperscript{2,3,4}

\textsuperscript{1} School of Biology, University of Leeds, LS2 9JT Leeds, UK.
\textsuperscript{2} Department of Collective Behaviour, Max Planck Institute for Ornithology, 78457 Konstanz, Germany.
\textsuperscript{3} Chair of Biodiversity and Collective Behaviour, Department of Biology, University of Konstanz, 78457 Konstanz, Germany.
\textsuperscript{4} Edward Grey Institute, Department of Zoology, University of Oxford, Oxford, OX1 3PS, UK.

Addresses for correspondence:
William Hoppitt: W.J.E.Hoppitt@leeds.ac.uk
Damien Farine: dfarine@orn.mpg.de

ABSTRACT

Social network analysis has provided important insight into many population processes in wild animals. Constructing social networks requires quantifying the relationship between each pair of individuals in the population. Researchers often use association indices to convert observations into a measure of propensity for individuals to be seen together. At its simplest, this measure is just the probability of observing both individuals together given that one has been seen (the simple ratio index). However, this probability becomes more challenging to calculate if the detection rate for individuals is imperfect. We first evaluate the performance of existing association indices at estimating true association rates under scenarios where (i) only a proportion of all groups are observed (group location errors), (ii) not all individuals are observed despite being present (individual location errors),
and (iii) a combination of the two. Commonly-used methods aimed at dealing with incomplete observations perform poorly because they are based on arbitrary observation probabilities. We then derive complete indices that can be calibrated for the different types of observation probabilities to generate accurate estimates of association rates. These are provided in an R package that readily interfaces with existing routines. We conclude that using calibration data is an important step when constructing animal social networks, and that in their absence, researchers should use a simple estimator and explicitly consider the impact of this on their findings.

KEYWORDS
Affiliations, Animal social networks, Interactions, Social network analysis, Social structure

INTRODUCTION
A foundation of animal social network analysis is estimating the frequency that two individuals associate or interact. Social networks are typically a description of interconnections that are formed by relationships (edges) among multiple individuals (nodes). Social network analysis is a set of tools that can be used to describe the patterns formed by these interconnections or evaluate these against hypotheses (Farine & Whitehead, 2015; Whitehead, 2008). One feature of social network analysis that is perhaps unique to studies on animal populations is that researchers rarely have a complete record of all interactions or all associations (but see Boogert, Farine, & Spencer, 2014; Farine, Spencer, & Boogert, 2015; Strandburg-Peshkin, Farine, Couzin, & Crofoot, 2015). Thus, relationships are often imperfectly sampled, which can introduce uncertainty in the social network. To account for variation in sampling effort and observation frequency, Cairns and Schwager (1987) outlined commonly-used association indices. These indices convert the number of observations of pairs of individuals seen associating or interaction into an...
association rate, representing their propensity to associate or their probability of being observed together.

Incomplete sampling of animal interactions or associations can occur due to a range of different reasons. We can classify datasets as having two possible types of missing data (Cairns & Schwager, 1987): (i) single or few observers can only collect data on one or a few groups at a time and miss many simultaneous associations or interactions occurring elsewhere, and (ii) individuals are difficult to observe or identify and missed even when they are present. In type (i), while a number of pairs of individuals (also known as dyads) are being observed together in one or more groups, the status of other individuals in the population is unobserved. In type (ii), when one or more groups are being observed they are incompletely sampled, resulting in data that suggests that certain dyads were not interacting or associating even when they were and could have been observed doing so. In both cases, the relationships inferred from the observed data is likely to be influenced by the amount of data that was missed. However, the propensity for each type of missing observations to impact our estimates of association or interaction rates and social network structure remains to be properly explored.

Properly controlling for missed observations is one of the most important steps in social network analysis. Using simulated data, Franks, Ruxton, and James (2010) identified the impact of missing observations when constructing social networks. They found that missing observations between known individuals was more problematic than missing individuals altogether, and concluded that social network sampling should maximize the amount of data collected about known individuals rather than maximizing the number of individuals sampled. One reason for this is because a key component of social networks, weak edges, are often disproportionately likely to be missed, and leaving these out can have profound implications on the structure of the social network (Granovetter, 1973). These findings are also supported by the work of M. J. Silk, Jackson, Croft, Colhoun, and Bearhop (2015) who explored the effect of completely missing individuals in the
social network. They found that, with adequate sampling, having as few as 30% of 
individuals known can be enough to produce informative networks for hypothesis 
testing.

Missing observations that could have been recorded can have large impacts on the 
social network that is generated, and these impacts are made worse when particular 
individuals are missed more often than others. Farine and Whitehead (2015) 
recently demonstrated how small differences in the likelihood of observing 
individuals of different classes can introduce systematic biases in their social 
network. They first simulated observations of individuals associating with preferred 
and avoided associates. They then introduced a small observation bias, in this case 
reducing the probability of observing one of two classes of individuals to 80% by 
removing 20% of the observations of those individuals. This resulted in a significant 
effect of class on degree (the sum of the association strengths in the nodes with 
intact data was higher than in the nodes where data had been removed). This means 
that the social network estimated for the individuals in this population is incorrect.

In this paper, we theoretically re-evaluate existing association indices and derive 
new measures to deal with missing observations of groups, missing individuals in 
groups, and the combination of these. We show that the extent that existing 
association indices adjust estimates of association strength is entirely arbitrary, and 
are as likely to over-correct any bias that might occur as they are to reduce it. 
Existing association indices can also perform poorly at estimating relative 
association strengths, which has implications for many social network studies. We 
them then derive improved association indices that enable researchers to correct 
properly for the biases arising from group location error and individual 
identification error, and discuss how to collect appropriate calibration data. Finally, 
we provide an R package “assocInd” that allows researchers to calculate accurate 
association indices for pairs of individuals from their observation data, and to 
simulate the effects of different types of errors on estimates of associations.
THE SIMPLE RATIO AND THE HALF-WEIGHT INDEX

In many cases we wish to calculate an association index that estimates the proportion of time any two individuals, a and b, spend associated. Association indices typically range from 0 (the two individuals were never observed together) to 1 (the individuals are always seen together), and the association rates are used as a proxy to quantify the propensity for pairs of individuals to interact (Farine, 2015; Whitehead & Dufault, 1999), although the assumption that individuals interact in proportion to their association rate should be considered on a case-by-case basis (Castles et al., 2014). Association data is frequently collected by repeatedly sampling the population, and recording whom is observed in the same group in each sampling period. For any two individuals we can then calculate:

- $x$ the number of sampling periods with a and b observed associated
- $y_a$ the number of sampling periods with just a identified
- $y_b$ the number of sampling periods with just b identified
- $y_{ab}$ the number of sampling periods with a and b identified but not associated
- $y_{null}$ the number of sampling periods with neither a nor b identified

In an ideal scenario, every individual is seen and correctly identified in every sampling period, such as in many captive populations, or at least we have the situation where $y_{null} = 0$. Intuitively, in the ideal scenario researchers can validly use the simple ratio index (SRI), $x/(y_a + y_b + y_{ab} + x)$, as an estimate of the proportion of time A and B spend together. However, when errors arise from missing observations of individuals or groups, it is less clear that the simple ratio is appropriate. The most commonly-used approach for correcting association indices to account for missing observations is to reduce the weighting given to observations of just one individual (because we have a lower confidence in these). Because missing observations are widespread in behavioural research, many researchers use the half-weight index (HWI): $x/(\gamma_y(y_a + y_b) + y_{ab} + x)$. This index is believed to
correct for the biases arising from such error, in particular when individuals are relatively more likely to be detected when they are apart than when they are together. When investigating the performance of association indices, Cairns and Schwager (1987) found that the HWI resulted in lower bias and lower error for a given estimate than the simple ratio when observations were missed. However, whilst this has served as useful justification for many researchers, it is also important to note that Cairns and Schwager (1987) reported up to 4 times greater error in the HWI than what they achieved using a maximum likelihood function (see also below). Further, they noted a number limitations of association indices arising from hidden assumptions.

Here we revisit some of the assumptions of the half-weight index. Notably, we show that the extent to which the half weight index adjusts estimates of association is entirely arbitrary, and is as likely to “overcorrect” any bias that might occur as it is to reduce that bias. Note that an alternative variant to the HWI, the twice-weight index \( (TWI) \frac{x}{2(y_a + y_b + y_{ab} + x)} \), is a monotonic function of the HWI and thus we do not investigate it in this paper. Ginsberg and Young (1992) previously raised the issue that the HWI and TWI use arbitrary weightings, and predicted that association indices will continue to be widely used. Indeed, the HWI is still the most commonly-used index in animal social network studies.

To address the need to properly correct for biases arising from group location error and individual identification error, we derive improved association indices that can be calibrated independently for each study. We start by addressing the impact of group location error before moving on to the effect of individual identification error, and finally the combination of the two.
CORRECTING FOR GROUP LOCATION ERROR

Here we start with the assumption that if you see a in a group but do not see b in that group, you know a is not with b, and vice versa. This assumption will be valid if there is no individual identification error, i.e. all the individuals in a group that is located by the researcher will always be identified. However, uncertainty remains for all of the sampling periods in which we did not see a or b, since we do not know if they were together during such periods.

Let us denote the event that a and b are together in a sampling period as \(ab\) and the event that they are not together as \(!ab\). The aim is to estimate the association between a and b, \(a_{ab} = p(ab)\). We start by developing a maximum likelihood estimator (MLE) for \(a_{ab}\). In any given sampling period, the probability we see only individual a (i.e. not b), is given by:

\[
P(\text{see } a, \text{not see } b) = p(\text{see } a| !ab) (1 - p(\text{see } b| !ab)) (1 - a_{ab})
\]

Note that \(p(!ab) = 1 - p(ab) = 1 - a_{ab}\). The probability of seeing a and b in different groups is given by:

\[
P(\text{see } a, \text{see } b) = p(\text{see } a| !ab) p(\text{see } b| !ab) (1 - a_{ab})
\]

The probability of seeing a and b together in a group is:

\[
P(\text{see } ab) = p(\text{see } ab|ab) a_{ab}
\]

And the probability of seeing neither a nor b is:

\[
P(\text{not see } a, \text{not see } b) = (1 - p(\text{see } a| !ab)) (1 - p(\text{see } b| !ab)) (1 - a_{ab}) + (1 - p(\text{not see } ab| ab)) a_{ab}
\]
From this we can derive the log-likelihood, $L$, for the data obtained:

$$ L = y_a \log(P(\text{see } a|ab)) + \log((1-P(\text{see } b|!ab))) + \log(1-a_{ab}) $$

$$ + y_b \log(P(\text{see } b|!ab)) + \log((1-P(\text{see } a|!ab))) + \log(1-a_{ab}) $$

$$ + y_{ab} \log(P(\text{see } a|!ab)) + \log(P(\text{see } b|!ab)) + \log(1-a_{ab}) $$

$$ + x \log(P(\text{see } ab|ab)) + \log(a_{ab}) $$

$$ + y_{\text{null}} \log((1-P(\text{see } a|!ab))(1-P(\text{see } b|!ab))(1-a_{ab}) + (1-P(\text{see } ab|ab))a_{ab}) $$

This simplifies to:

$$ L = (y_a + y_b + y_{ab}) \log(1-a_{ab}) + x \log(a_{ab}) + y_a \log(1-A) + \log(B) + y_b \log(1-B) + \log(A) + y_{ab} \log(1-A) + \log(1-B) + x \log(1-C) + y_{\text{null}} \log[AB(1-a_{ab}) + C a_{ab}] $$

where $A = P(\text{not seeing } a \text{ when } a \text{ and } b \text{ are not together}), B = P(\text{not seeing } b \text{ when } a \text{ and } b \text{ are not together}),$ and $C = P(\text{not seeing } a \text{ or } b \text{ when they are together})$ will not be known. Thus, to estimate an accurate value for the association strength between two individuals requires validation data at the level of individuals. However, progress can be made by making different assumptions about the relationship between $A, B$ and $C$. First, if $C = AB$, the MLE is:

$$ \hat{a}_{ab} = \frac{x}{(y_a + y_b + y_{ab} + x)} $$

which is the simple ratio index. So, if the probability of failing to see $a$ and $b$ together is the same as the probability of failing to see both when they are apart, then the SRI
is valid. Note also that the simple ratio is valid as an MLE if \( y_{\text{Null}} = 0 \), as intuition suggests.

Alternatively, we could assume that \( C = AB(1 + \omega) \), where failing to observe a group containing \( a \) and \( b \) is more (\( \omega > 0 \)) or less (\( \omega < 0 \)) likely than failing to observe both the group containing \( a \) and the group containing \( b \) when \( a \) and \( b \) are not together. In this case the MLE is given by the solution to:

\[
\frac{(y_a + y_b + y_{ab})}{1 - \hat{a}_{ab}} + \frac{x}{\hat{a}_{ab}} + y_{\text{Null}} \frac{\omega}{\hat{a}_{ab} + 1} = 0
\]

which can be re-arranged to form a quadratic:

\[
-\omega T \hat{a}_{ab}^2 + (x\omega + y_{\text{Null}}\omega - T_i)\hat{a}_{ab} + x = 0,
\]

where \( T_i \) is the number of directly informative sampling periods, i.e. \( T_i = y_a + y_b + y_{ab} + x \), and \( T \) is the total number of sampling periods, i.e. \( T = I + y_{\text{Null}} \). The MLE is given by the lower root of this equation, i.e.

\[
\hat{a}_{ab} = \frac{-(x\omega + y_{\text{Null}}\omega - T_i) - \sqrt{(x\omega + y_{\text{Null}}\omega - T_i)^2 + 4\omega T x}}{2\omega T}
\]

We term this index the group location error corrected index (GLECI). As expected, the GLECI reduces to the simple ratio index \( \hat{a}_{ab} = x/T \) when \( y_{\text{Null}} = 0 \). The standard error (see Appendix for derivation) can be calculated as for a proportion, \( \sqrt{a_{ab}(1 - a_{ab})/N_e} \), but with an effective sample size of:

\[
N_e = T_i + y_{\text{Null}} \frac{\omega a_{ab}(1 - a_{ab})}{(\omega a_{ab} + 1)^2}
\]
assuming sampling periods are sufficiently spaced in time and/or space to assume
they are independent.

CORRECTING FOR INDIVIDUAL IDENTIFICATION ERROR

Here we assume that there is no group location error, but define the probability of
failing to identify an individual in a group that has been under observation, i.e. the
individual observation error rate, as $\epsilon$. We suspect this scenario will be rare, as
individual identification error will usually be accompanied by group location error
(see next section). However, we consider the case in order to analyse the effect of
the two types of error. We get the following probabilities:

\[
\begin{align*}
    p(\text{see } a, \text{see } b) &= (1 - \epsilon)^2 (1 - a_{ab}) \\
    p(\text{see } ab) &= (1 - \epsilon)^2 a_{ab} \\
    p(\text{not see } a, \text{not see } b) &= \epsilon^2 \\
    p(\text{see } a, \text{not see } b) &= +\epsilon(1 - \epsilon)
\end{align*}
\]

Therefore the likelihood for the data can be obtained as:

\[
L = y_{ab} [2 \log(1 - \epsilon) + \log(1 - a_{ab})] + x [2 \log(1 - \epsilon) + \log(a_{ab})] + y_{null} [2 \log(\epsilon)] + (y_a + y_b) [\log(\epsilon) + \log(1 - \epsilon)]
\]

And the MLE found as follows:

\[
\frac{dL}{da_{ab}} = -y_{ab} \frac{1}{1 - a_{ab}} + x \frac{x}{a_{ab}} = 0
\]

\[
(-y_{ab} - x)a_{ab} + x = 0
\]

\[
a_{ab} = \frac{x}{y_{ab} + x}
\]
In this case the MLE $\widehat{a_{ab}}$ is a simple proportion, using only the $x$ and $y_{ab}$ counts taken to the informative data, and requiring no calibration data. The terms $y_a$ and $y_b$ are not used, in contrast to the simple ratio and half weight indices, since they are known to be unreliable in this scenario: if only $a$ is recorded, it is possible that $b$ was associated with $a$ and has been missed through individual identification error. We call this index the very simple ratio index (vSR). The standard error (see Appendix for derivation) is calculated for a proportion as usual with $\sqrt{a_{ab}(1-a_{ab})/N_e}$, with an effective sample size of $N_e = y_{ab} + x$, assuming sampling periods are sufficiently spaced in time and/or space to assume they are independent.

GENERAL ERROR MODEL

In practice, both types of error are likely to occur in a given sampling procedure. To model this situation, we define a model with a more general relationship between the errors in each count. We define:

$$p(\text{see } a, \text{see } b) = (1 - \epsilon_{a|ab})(1 - \epsilon_{b|ab})(1 - a_{ab})$$

$$p(\text{see } ab) = (1 - \epsilon_{a|ab} - \epsilon_{b|ab} - \epsilon_{ab|ab})a_{ab}$$

$$p(\text{not see } a, \text{not see } b) = \epsilon_{ab|ab}a_{ab} + \epsilon_{a|ab} \epsilon_{b|ab}(1 - a_{ab})$$

$$p(\text{see } a, \text{not see } b) = \epsilon_{b|ab}a_{ab} + (1 - \epsilon_{a|ab}) \epsilon_{b|ab}(1 - a_{ab})$$

where $\epsilon_{a|ab}$ is the probability of missing $a$, given $a$ is not with $b$; $\epsilon_{a|ab}$ is the probability of missing $a$, but not $b$, given $a$ is with $b$; and $\epsilon_{ab|ab}$ is the probability of missing $a$ and $b$ given they are together.

Note that the group location error scenario is the special case with $\epsilon_{a|ab} = A$;

$\epsilon_{b|ab} = B; \epsilon_{ab|ab} = C = (1 + \omega)AB = (1 + \omega) \epsilon_{a|ab} \epsilon_{b|ab}; \epsilon_{a|ab} = \epsilon_{b|ab} = 0$. The
individual identification error scenario is given by $\varepsilon_{a|ab} = \varepsilon_{b|ab} = \varepsilon$; $\varepsilon_{a|ab} = \varepsilon_{b|ab} = \varepsilon(1 - \varepsilon)$; $\varepsilon_{ab|ab} = \varepsilon^2$.

Let us assume that the probability of missing both $a$ and $b$ when they are in the same group is $(1 + \omega) \times$ the probability of missing $a$ and $b$ when they are not together (where $\omega > -1$). The probability both $a$ and $b$ will be missed when they are together is therefore:

$$\varepsilon_{ab|ab} = (1 + \omega) \varepsilon_{a|ab} \varepsilon_{b|ab}$$

The probability of at least one of $a$ or $b$ being missed is $\hat{\varepsilon} = 1 - p(\text{see } ab|ab)$. Let us set $\varepsilon_{ab|ab} = \phi \hat{\varepsilon}$, where $0 < \phi \leq 1$. Since $\hat{\varepsilon} = \varepsilon_{a|ab} + \varepsilon_{b|ab} + \varepsilon_{ab|ab}$

$$\varepsilon_{a|ab} + \varepsilon_{b|ab} = (1 - \phi) \hat{\varepsilon}$$

$$\varepsilon_{a|ab} + \varepsilon_{b|ab} = \frac{(1 - \phi)(1 + \omega) \varepsilon_{a|ab} \varepsilon_{b|ab}}{\phi}$$

If we assume that $\varepsilon_{a|ab} / (\varepsilon_{a|ab} + \varepsilon_{b|ab}) = \varepsilon_{a|ab} / (\varepsilon_{a|ab} + \varepsilon_{b|ab})$, this gives us:

$$\varepsilon_{a|ab} = \frac{(1 - \phi)(1 + \omega) \varepsilon_{a|ab}^2 \varepsilon_{b|ab}}{\phi(\varepsilon_{a|ab} + \varepsilon_{b|ab})}$$

where $\phi$ determines the relative importance of group location error relative to individual identification error, with the group location error model given when $\phi = 1$. We can now refine the probabilities given above:

$$p(\text{see } a, \text{see } b) = (1 - \varepsilon_{a|ab})(1 - \varepsilon_{b|ab})(1 - a_{ab})$$

$$p(\text{see } ab) = (1 - (1 + \omega) \varepsilon_{a|ab} \varepsilon_{b|ab} / \phi)a_{ab}$$

$$p(\text{not see } a, \text{not see } b) = (1 + \omega) \varepsilon_{a|ab} \varepsilon_{b|ab} a_{ab} + \varepsilon_{a|ab} \varepsilon_{b|ab} (1 - a_{ab})$$
\[
p(\text{see a, not see b}) = \varepsilon_{b|ab} \left( (1 - \phi)(1 + \omega) \frac{\varepsilon_{b|ab} \varepsilon_{a|ab}}{\phi(\varepsilon_{a|ab} + \varepsilon_{b|ab})} a_{ab} + (1 - \varepsilon_{a|ab})(1 - a_{ab}) \right) + \varepsilon_{a|ab} (1 - \varepsilon_{b|ab})(1 - a_{ab})
\]

Giving a log likelihood of:

\[
L = y_a \left[ \log(\varepsilon_{b|ab}) + \log\left( \frac{(1 - \phi)(1 + \omega) \varepsilon_{b|ab} \varepsilon_{a|ab}}{\phi(\varepsilon_{a|ab} + \varepsilon_{b|ab})} a_{ab} + (1 - \varepsilon_{a|ab})(1 - a_{ab}) \right) \right] + y_b \left[ \log(\varepsilon_{a|ab}) + \log\left( \frac{(1 - \phi)(1 + \omega) \varepsilon_{a|ab} \varepsilon_{b|ab}}{\phi(\varepsilon_{a|ab} + \varepsilon_{b|ab})} a_{ab} + (1 - \varepsilon_{b|ab})(1 - a_{ab}) \right) \right] + y_{ab} \left[ \log(1 - \varepsilon_{a|ab}) + \log(1 - \varepsilon_{b|ab}) + \log(1 - a_{ab}) \right] + y_{null} \left[ \log\left( (1 + \omega) \varepsilon_{a|ab} \varepsilon_{b|ab} a_{ab} + \varepsilon_{a|ab} \varepsilon_{b|ab} (1 - a_{ab}) \right) \right]
\]

To obtain the MLE, \( \hat{a}_{ab} \), we need to solve the equation:

\[
\frac{dL}{da_{ab}} = y_a \left( (1 - \phi)(1 + \omega) \frac{\varepsilon_{b|ab} \varepsilon_{a|ab}}{\phi(\varepsilon_{a|ab} + \varepsilon_{b|ab})} a_{ab} + (1 - \varepsilon_{a|ab})(1 - a_{ab}) \right) - y_b \left( (1 - \phi)(1 + \omega) \frac{\varepsilon_{a|ab} \varepsilon_{b|ab}}{\phi(\varepsilon_{a|ab} + \varepsilon_{b|ab})} a_{ab} + (1 - \varepsilon_{b|ab})(1 - a_{ab}) \right) - y_{ab} \frac{1 - \hat{a}_{ab}}{1 - \hat{a}_{ab}}
\]

\[
+ y_{null} \left( (1 + \omega) \varepsilon_{a|ab} \varepsilon_{b|ab} \varepsilon_{a|ab} \varepsilon_{b|ab} a_{ab} + \varepsilon_{a|ab} \varepsilon_{b|ab} (1 - a_{ab}) \right) = 0
\]
To generate an estimate of association using this function firstly requires values for \( \omega \) (the group location error term) and \( \phi \) (the error components importance term). These could reasonably be calibrated at the population level, i.e. we could assume that these quantities are constant across all dyads. However, the estimate also requires estimates for \( \varepsilon_{a|ab} \) and \( \varepsilon_{a|lab} \) which require calibration data at the level of individuals, making this approach infeasible in most cases. Nonetheless, we might obtain an approximate solution, \( \widetilde{a}_{ab} \), if we substitute a population averaged estimate (averaged across all dyads, or those dyads for which data is available) \( \varepsilon = \varepsilon_{a|lab} = \varepsilon_{b|lab} \):

\[
(\gamma_a + \gamma_b) \frac{((1 - \phi)/\phi)(1 + \omega) \in -2(1-\varepsilon)}{((1 - \phi)/\phi)(1 + \omega) \in \widetilde{a}_{ab} + 2(1-\varepsilon)(1 - \widetilde{a}_{ab})} - \frac{y_{ab}}{1 - \widetilde{a}_{ab}} + \frac{x}{\widetilde{a}_{ab}} + \frac{y_{null}(\omega)}{\omega \widetilde{a}_{ab} + 1} = 0
\]

The calibration measures required to solve this equation are \( \omega, \phi \) and \( \varepsilon \), and can be solved using a non-linear equation solver. We call the solution to this equation, \( \widetilde{a}_{ab} \), the combined errors index (CEI). In our R package, we provide a function that calculates the CEI in the R statistical environment (R Development Core Team, 2015), using the uniroot function in the rootSolve package (Soetaert & Herman, 2009). Note that by setting \( \phi = 1 \), we reduce the model to the group location error model, giving \( \widetilde{a}_{ab} \) as the GLECI. We can also reduce the model to the individual identification error model by setting \( \varepsilon_{a|lab} = \varepsilon_{b|lab} = \varepsilon (1-\varepsilon) \) and \( \varepsilon_{ab|lab} = \varepsilon^2 \), giving us \( \phi = \varepsilon/(2-\varepsilon) \) and \( \omega = 0 \). Thus \( \widetilde{a}_{ab} \) reduces to the vSRI.

**COMPARISON OF INDEX PERFORMANCE**

In this section we examine how the SRI, HWI, GLECI, vSRI and CEI perform under scenarios where there is group location error, individual identification error, and a
combination of the two. In simple cases we do this by first deriving expressions for
the expected value of each index, and then dividing by the value it is intended to
estimate, \(E[I_{ab}]/a_{ab}\), thus showing us the circumstances under which \(I_{ab}\) is biased
upwards or downwards. However, we also recognise that in many circumstances
only the relative sizes of \(a_{ab}\) within the social network may be required, e.g. if
estimating scale free node-based or network metrics. Consequently, we also derive
\((E[I_{ab}]/E[I_{uv}])/(a_{ab}/a_{uv})\) to determine the circumstances under which each index
tends to overestimate or underestimate ratios of association values. Here \(u\) and \(v\)
denote a different dyad, so \((E[I_{ab}]/E[I_{uv}])/(a_{ab}/a_{uv})\) measures the bias when
index \(I_{ab}\) is used to estimate the relative strength of two associations.

In each case we also use simulations to illustrate the performance of the indices for
a given set of values, investigate bias where we were unable to do so analytically,
and examine the performance of Wald 95\% confidence intervals calculated from the
standard errors presented above. For each scenario, we simulated 10,000 datasets
consisting of 1000 independent sampling periods for two individuals \(a\) and \(b\), in the
R statistical environment (R Development Core Team, 2015). We start by allocating
the probability \(a_{ab} = 0.5\) that \(a\) and \(b\) were associating in a given sampling period.
We then repeated all simulations with \(a_{ab} = 0.25\) and \(a_{ab} = 0.75\), and the results
we qualitatively similar, so here we present the results for \(a_{ab} = 0.5\) only. We then
simulated the observation process according to the models described above, to yield
values for \(y_a, y_b, y_{ab}, x\) and \(y_{null}\) which we used to calculate the value of each of the
target association indices. For each scenario, we repeated the simulation for a range
of values of group location and individual identification errors. For the scenario
including only group location error, we ran simulations for a range of values of
\(\omega = \{-0.9, -0.8, \ldots, 2.5\}\) with \(A = B = 0.5\), where \(A = P(\text{see } a|ab)\) and \(B = P(\text{see } b|\neg ab)\). For the scenario including only individual identification error, we ran
simulations for a range of values of \(\epsilon = \{0.05, \ldots, 0.95\}\). For the scenario with both
types of error, we varied \(\omega = \{-0.9, -0.8, \ldots, 0.9\}\), \(\epsilon = \{0.1, 0.3, 0.5\}\) and
\(\phi = \{0.1, 0.3, 0.5\}\), excluding impossible cases where \((1 + \omega)^2/\phi > 1\), since this
would mean there is a negative probability of observing \(a\) and \(b\) together. In each
case we recorded the mean value for each association index, in order to detect bias, and the proportion of times the Wald 95% confidence intervals (calculated as $\pm 1.96 \times \text{SE}$) contain the true value for $a_{ab}$.

However, recall that the CECI relies on an approximation, $\epsilon = \epsilon_{a|ab} = \epsilon_{b|ab}$, which replaces the individual specific error rates with population level ones. The simulations described above only test the performance of the CE when this approximation holds in the data, i.e. when error rates are the same across all individuals. Consequently, we re-ran simulations to test the performance of the CECI when there was individual variation in error rate. In each case we set the population mean error, $\epsilon$, arbitrarily to 0.5, but drew individual errors from a normal distribution with standard deviation $\sigma = \{0,0.2,\ldots,2.0\}$, discarding and resampling values that were $< 0$ or $> 1$, and likewise for $\epsilon_{b|ab}$. We then conducted the simulations as described above with $\phi = 0.5$.

**Group location error only**

We find that the GLECI is an unbiased estimator of $a_{ab}$ (see Table 1 and Fig. 1) across a range of group location errors $\omega$. By contrast, the simple ratio is biased upwards when $\omega < 0$ (i.e. when $a$ and $b$ are less likely to be missed when associated than both are to be missed when apart) and biased downwards when $\omega > 0$. The commonly used HWI shifts the estimate of $a_{ab}$ upwards, such that it is biased upwards when $AB(1 + \omega) < (A + B)/2$ and biased downwards when $AB(1 + \omega) > (A + B)/2$. Consequently the HWI is only unbiased when the probability of seeing $a$ and $b$ together is equal to the average of the probability of seeing each of them apart. In our terminology, this is denoted $(1 - C) = ((1 - A) + (1 - B))/2$, giving $C = (A + B)/2$. Thus under this scenario, the HWI assumes that the probability of missing $a$ and $b$ when they are together is equal to half the probability of missing either $a$ or $b$ when they are apart. This seems to us to be an arbitrary a priori assumption, without the functionality to adjust the assumption using supporting calibration data.
The vSRI performs very poorly when only group location error is present, overestimating $a_{ab}$ whilst $\omega < (A + B - 2AB)/AB$. This is because it excludes $y_a$ and $y_b$ from the denominator on the assumption that these data are uninformative, whereas in this scenario these are cases where we know that $a$ and $b$ were not associating. As $\omega$ increases, an increasing number of cases where $a$ and $b$ were associating are erroneously assigned to $y_{null}$. Thus, exclusion of $y_{null}$ from the index eventually offsets the positive bias (when $\omega = (A + B - 2AB)/AB$) resulting from exclusion of $y_a$ and $y_b$ from the denominator. The vSRI is therefore not a useful index as it contains assumptions that are unlikely to be met in the majority of studies.

The GLECI is generally unbiased because $y_{null}$ is included in the index in such a way that excludes this positive bias. Importantly, the 95% confidence intervals for the GLECI contained the true value of $a_{ab}$ in close to 95% of cases, showing they perform validly in this scenario (see Fig 1b). In contrast, the 95% confidence intervals associated with the simple ratio, half weight index and IIIECI index only performed acceptably for a very narrow range of values of $\omega$. Furthermore, the GLECI (and CEI index with $\phi = 1$) is the only index of those considered that is unbiased when estimating the ratio of two associations (see Table 1). This suggests that the SRI, HWI and IIIECI are not suitable for estimating either the relative or absolute strength of associations when group location error is believed to be present, and that, ideally the GLECI should be used if calibration data can be obtained.

Note that the simulations assumed that the researcher has an accurate estimate of $\omega$ with which to calculate the GLECI (see below). In reality, the better the estimate of $\omega$ is, the better the estimate of $a_{ab}$ will be, but even a rough estimate of $\omega$ will be preferable to no calibration at all. Furthermore, we assume that the probability of missing $a$ and $b$ when they are together will be $(1 + \omega) \times$ that of missing both $a$ and $b$ when they are separate ($C = (1 + \omega)AB$). Further work may conclude that this relationship does not generally hold, in which case the GLECI might be suitably modified to use a different calibration statistic. Nonetheless, the relationship posited here requires weaker a priori assumptions to be made about the data than the
commonly used half weight index, which assumes that $C = (A+B)/2$. We show that when this assumption is even slightly wrong, the half weight index will be a poor estimate of $a_{ab}$.

One possible option to resolve the half weight index is to generalize it to be an M weight index (MWI): $x/(m(y_a + y_b + y_{ab} + x))$. The M weight index assumes that the probability of seeing $a$ and $b$ together is equal to $M_x$ of the sum of the probability of seeing each of them each apart. Equivalently, missing $a$ and $b$ when they are together is $M_x$ more likely than missing them both when they are apart. Thus the MWI could be calibrated to the data analogously to the GLECI, but each index assumes a different relationship among the observation errors in the population.

When calibration data cannot be obtained, there is a strong case for preferring the simple ratio index. Use of the SRI results in biases that are more likely to be qualitatively predictable in the absence of calibration data, than biases resulting from the HWI or IIECI (see Table 1). For instance, when a researcher suspects that, in general, missing two individuals when they are together is more likely than missing both when they are apart, they can expect i) SRI values to be underestimates of $a_{ab}$; ii) for this underestimation to be more pronounced for smaller values of $a_{ab}$ and for less commonly seen individuals; and iii) for bigger ratios between real associations to be overestimated relative to smaller ratios. Thus a researcher can assess whether these inaccuracies are likely have any great bearing on their conclusions in their specific case. In contrast, using the HWI we cannot easily make such qualitative predictions unless we are in a position to judge the relative size of $AB(1 + \omega)$ versus $(A + B)/2$; and, for the IIECI, $\omega$ versus $(A + B - 2AB)/AB$. Making such judgments empirically is likely to be at least as challenging as acquiring the calibration data required calculating the GLECI or a calibrated MWI. Consequently, if only group location error is present and calibration data cannot be obtained, we recommend use of the SRI, with careful consideration of how the biases identified above might affect the interpretation of the study.
Table 1: Biases in different association measures arising from group location error.

*signifies no bias inferred from simulations.

<table>
<thead>
<tr>
<th>Measure</th>
<th>$I_{ab}$</th>
<th>$E[I_{ab}] / a_{ab}$</th>
<th>$E[I_{ab} / I_{uv}] / (a_{ab} / a_{uv})$</th>
</tr>
</thead>
</table>
| SRI     | \[
\frac{(1 - AB(1 + \omega))}{1 - AB(1 + a_{ab}\omega)}
\] | \[
\frac{(1 - AB(1 + \omega))(1 - CD(1 + a_{cd}\omega))}{(1 - CD(1 + \omega))(1 - AB(1 + a_{ab}\omega))}
\] |
| HWI     | \[
\frac{(1 - AB(1 + \omega))(1 - a_{ab})(1 - (A + B)/2) + (1 - AB(1 + \omega))a_{ab}}{(1 - a_{ab})(1 - (A + B)/2) + (1 - AB(1 + \omega))a_{ab}}
\] | \[
\frac{(1 - AB(1 + \omega))}{(1 - a_{uv})(1 - (U + V)/2) + (1 - UV(1 + \omega))a_{uv}}
\] |
| IIECI   | \[
\frac{(1 - AB(1 + \omega))(1 - A)(1 - B) + a_{ab}(A + B - 2AB - AB\omega)}{(1 - A)(1 - B) + a_{ab}(A + B - 2AB - AB\omega)}
\] | \[
\frac{(1 - AB(1 + \omega))(1 - (A + B - 2AB - AB\omega))(1 - U)(1 - V)}{(1 - (A + B - 2AB - AB\omega))(1 - a_{uv}(U + V - 2UV - UV\omega)}
\] |

- **Unbiased when** $\omega = 0$
- **Underestimates** $a_{ab}$ when $\omega > 0$: effect more pronounced for smaller values of $a_{ab}$ and less commonly seen individuals
- **Overestimates** $a_{ab}$ when $\omega < 0$: effect more pronounced for smaller values of $a_{ab}$ and less commonly seen individuals
- **When** $\omega > 0$ bigger ratios exaggerated and biased in favour of the $a_{ab}$, for pairs of individuals more commonly seen
- **When** $\omega < 0$ smaller ratios exaggerated and biased in favour of the $a_{ab}$, for pairs of individuals less commonly seen
- **Unbiased when** $AB(1 + \omega) = (A + B)/2$
- **Otherwise difficult to predict the pattern of bias in the data**
- **Unbiased when** $\omega = (A + B - 2AB)/AB$
\[
\omega > (A + B - 2AB) / AB \\
\cdot \text{Overestimates } a_{xy} \text{ when } \omega < (A + B - 2AB) / AB
\]

<table>
<thead>
<tr>
<th>GLECI /</th>
<th>1</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>CECI</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- Unbiased for this error model*
- Unbiased for this error model*

**Figure 1:** a) Bias in different association indices as a function of group location error (\(\omega\)) when applied to simulated data; b) performance of 95% Wald confidence intervals as a function of \(\omega\). Similar results were obtained for a true association value of 0.25 and 0.75.
Individual identification error only

When only individual identification error is present we find that the IIECI is an unbiased estimator of $a_{ab}$, whereas the simple ratio and half weight indices are biased downwards whenever $\epsilon > 0$ (see Table 2 and Figure 2). This is because as $\epsilon$ increases, more and more cases where $a$ and $b$ are associating are erroneously attributed to $y_a$ or $y_b$. Since $y_a$ and $y_b$ are included in the denominator for the SRI and HWI, this results in an under-estimation of $a_{ab}$. The effect is reduced in the HWI since $y_a$ and $y_b$ have a reduced weighting in this index. However, this is not sufficient to ensure that the HWI is valid for even small individual identification error rates. The GLECI was not included separately in these simulations since in this scenario $(1 + \omega)e^2 = \epsilon^2$, so $\omega = 0$, meaning the GLECI reduces to the simple ratio.

Consequently, we can see that the GLECI also performs badly at estimating absolute association values when there is individual identification error but no group location error.

The 95% confidence intervals for the vSRI contained the true value of $a_{ab}$ in close to 95% of cases, showing they perform validly under a scenario of only individual identification error (see Fig 2b). This dropped slightly with very high error rates, as a result of the effective sample size decreasing as most data is attributed to $y_a, y_b$ or $y_{null}$ (not that with small sample sizes Wald confidence intervals are likely to be anti-conservative: too narrow). In contrast, the 95% confidence intervals associated with the SRI (and hence the GLECI) and HWI performed very badly with even small individual identification error rates. Consequently, if a researcher is interested in estimating the absolute values of $a_{ab}$ and only individual identification error is likely to be present (we anticipate this scenario to be rare), we recommend use of the vSRI, which requires no calibration data.

In contrast to the group location error only, all association indices gave unbiased estimates of the relative size of associations between pairs of individuals (see Table 2). This means that if the research aims are purely in the scale free properties of a system, such as the relative position of individuals in a social network (Aplin, Firth,
et al., 2015; Wilson, Krause, Dingemanse, & Krause, 2013), any of the indices considered will be sufficient. However, if there is also a risk of group location error, and appropriate calibration data cannot be obtained (see next section), we recommend use of the SRI to estimate relative associations due to the advantages in interpreting this index in the presence of group location error (see above).

Figure 2: a) Bias in different association indices as a function of individual identification error ($\omega$) when applied to simulated data; b) performance of 95% Wald confidence intervals as a function of $\omega$. Similar results were obtained for a true association value of 0.25 and 0.75.
Table 2: Biases in different association measures arising from individual identification error

<table>
<thead>
<tr>
<th>$I_{ab}$</th>
<th>$E[I_{ab}]/a_{ab}$</th>
<th>$E[I_{ab}/I_{uw}]/(a_{ab}/a_{uv})$</th>
</tr>
</thead>
<tbody>
<tr>
<td>SRI</td>
<td>$1 + \varepsilon(2 - \varepsilon)$</td>
<td>1</td>
</tr>
<tr>
<td>HWI</td>
<td>$1 + \varepsilon$</td>
<td>1</td>
</tr>
<tr>
<td>IIECI/CECI</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>GLECI</td>
<td>$1 + \varepsilon(2 - \varepsilon)$</td>
<td>1</td>
</tr>
</tbody>
</table>

Combined errors

We find that the CECI is an unbiased estimator of $a_{ab}$ across the possible range of values for $\varepsilon$, $\omega$ and $\phi$ when we assumed $\varepsilon_{a|ab} = \varepsilon$ for all individuals (see Fig S1 in ESM). Furthermore, the CECI was also an unbiased estimator of $a_{ab}$ when $\varepsilon_{a|ab}$ was allowed to vary across the population regardless of the magnitude of variation in $\varepsilon_{a|ab}$ (see Fig S2 in ESM). In contrast, the SRI, HWI, IIECI and GLECI were biased in a manner that was dependent on the combination of values for $\varepsilon$, $\omega$ and $\phi$. When $\phi$ was close to 1, the pattern of bias was similar to when only group location error was present. In other words, when it is unlikely that only one of $a$ and $b$ will be missed when they are in the same group, bias is similar to when we have only group location error. As $\phi$ became smaller (more likely that only one of $a$ and $b$ will be missed when they are in the same group) all four indices start to underestimate $a_{ab}$ at a lower value of $\omega$. The effects of both $\phi$ and $\omega$ are magnified more as $\varepsilon$ gets larger. Consequently, if individual identification errors are likely to be common in addition to group location error, we suggest calibration data is acquired to estimate $\varepsilon$, $\omega$ and $\phi$ and the CECI is used. The pattern of bias in the other indices will be difficult to predict qualitatively unless the risk of individual identification error is known to be small. Consequently, if calibration data cannot be obtained under such circumstances we suggest extra efforts are made to minimise individual
identification error, and the SRI be used with the understanding that it will provide noisy estimates of $a_{ab}$.

When we assume observation error is homogeneous across the population, the 95% confidence intervals for the CECI tend to contain the true value of $a_{ab}$ in >95% of cases (see Fig S3). This suggests that, at large sample sizes at least, the Wald confidence intervals are slightly too wide. However, given that Wald confidence intervals are always an approximation and are widely used in statistics, this is a minor concern. However, when observation error varied greatly across the population, the 95% confidence intervals became far too narrow (see Fig S4), as a result of the extra uncertainty that is unaccounted for in the derivation of the standard error. Correcting the standard errors for this uncertainty does not seem straightforward, though further work could address this if the CECI proves to be useful and becomes widely adopted. Therefore, we suggest that the standard errors and confidence intervals for the CECI be trusted as approximately valid if the variation in observation rate is believed to be small, and not be trusted if that variation is believed to be large.

Our recommendations for the choice of association index are shown as a flowchart in Fig. 3. The indices and their standard errors are shown in Table 3.

**OBTAINING CALIBRATION DATA**

Here we suggest some initial ideas for obtaining calibration data that can be used to estimate the calibration parameters derived above. These suggestions can almost certainly be improved upon, by taking account of what data collection protocols are feasible in specific circumstances perhaps by deriving maximum likelihood estimates of calibration parameters given the data yielded by each such protocol. Here we limit ourselves to providing relatively simple intuitive ways of estimating calibration parameters. Whilst not optimal, our analysis above suggests that these methods are nonetheless likely to be an improvement on the unsupported use of a non-calibrated index such as the HWI.
Table 3: Summary of the indices considered in this paper.*Standard errors presented assume sampling periods are sufficiently spaced to be considered approximately independent, and that any calibration measures are known without error. See Supplementary Information for derivations of standard errors.

<table>
<thead>
<tr>
<th>Index name</th>
<th>Formula for $I_{ab}$</th>
<th>Standard error*</th>
<th>Effective sample size ($N_e$)</th>
<th>Calibration measures required</th>
</tr>
</thead>
<tbody>
<tr>
<td>Simple ratio index (SRI)</td>
<td>$x/(y_a + y_b + y_{ab} + x)$</td>
<td>$y_a + y_b + y_{ab} + x$</td>
<td>None</td>
<td></td>
</tr>
<tr>
<td>Half weight index (HWI)</td>
<td>$x/\left(\frac{2}{3}(y_a + y_b) + y_{ab} + x\right)$</td>
<td>$y_a + y_b + y_{ab} + x$</td>
<td>None</td>
<td></td>
</tr>
<tr>
<td>M weight index (MRI)</td>
<td>$x/(M(y_a + y_b) + y_{ab} + x)$</td>
<td>$y_a + y_b + y_{ab} + x$</td>
<td>M</td>
<td></td>
</tr>
<tr>
<td>Group location error</td>
<td>$(\sqrt{(x \omega + y_{null} \omega - T_i)^2 + 4 \omega I x}^-1)_{Ne}$</td>
<td>$y_a + y_b + y_{ab} + x$</td>
<td>$\omega$</td>
<td></td>
</tr>
<tr>
<td>corrected index (GLECI)</td>
<td>$\omega I_{ab} (1 - I_{ab})$</td>
<td>$y_a + y_b + y_{ab} + x$</td>
<td>$\omega$</td>
<td></td>
</tr>
<tr>
<td>Very simple ratio index</td>
<td>$x/(y_{ab} + x)$</td>
<td>$y_{ab} + x$</td>
<td>None</td>
<td></td>
</tr>
<tr>
<td>(vSRI)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Combined error corrected</td>
<td>Not available in closed form (see assocInd package for R code to generate this index and standard error)</td>
<td>$\epsilon, \omega$ and $\phi$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>index (CECI)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
One way a researcher might obtain estimates of calibration parameters is to collect data that can be assumed to be approximately error free for a subset of individuals, whilst simultaneously collecting association data using their standard protocol. This could be done by focal follows of a sample of individuals conducted by one researcher, whilst another collects data using the association data collection protocol. Alternatively some individuals could be tagged with GPS or proximity loggers (Kays, Crofoot, Jetz, & Wikelski, 2015; Krause et al., 2013) able to record encounters between individuals with more precision.

First we suggest that a researcher assess whether or not individual identification error is present and important. This could be done by calculating the proportion of sampling periods in which each individual included in the error free dataset was

---

*Validity of SE assumes sampling periods are sufficiently spaced*

---

**Figure 3**: Flowchart with our suggested strategy for selecting an association index.
present in a group that was located using the standard protocol but not recorded as being present. If this proportion is 0 or close to 0 for most individuals, we suggest individual identification error be ignored, and the GLECI or MWI can be used. Otherwise, the CECI should be used (unless group location error is believed not to be present, in which case the vSRI should be used, which does not require calibration).

If individual identification error is not present or negligible, the researcher needs to choose between the GLECI and MWI and then estimate the relevant calibration parameter (\( \omega \) or \( m \)). For any two individuals \( a \) and \( b \) in the error free sample, we know in which sampling periods they were together and in which they were not together. This enables the researcher to calculate \( y_{null|ab} \), the number of times \( a \) and \( b \) were not recorded by the association protocol during the calibration data collection period, when \( a \) and \( b \) were known to be together. \( C \), the probability \( a \) and \( b \) will be missed when they are together, can be estimated as \( y_{null|ab} / N_{ab} \) where \( N_{ab} \) is the number of sampling periods that \( a \) and \( b \) were known to be together. \( A \) can then be estimated as the proportion of sampling periods in which \( a \) was not recorded by the association protocol and known not to be with \( b \). \( B \) can be estimated in an analogous manner.

This process can be repeated for every combination of two individuals in the error free sample. The researcher can then use plots of these data to choose between the GLECI and MWI. If the assumptions of the GLECI hold, we would expect \( C \) to have a linear relationship with \( AB \), with a slope of \( (1 + \omega) \), whereas if the assumptions of the MWI hold, we would expect \( C \) to have a linear relationship with \( (A+B) \) with a slope of \( 1/m \). We suggest the researcher make each plot to decide which assumption is most realistic, and thus chose between the GLECI and MWI. If the GLECI is chosen, they can then fit a linear regression (constrained to pass through the origin) of \( C \) against \( AB \) and take \( \omega = \text{slope} - 1 \). If the MWI is chosen, they can then fit a linear
regression (constrained to pass through the origin) of \( C \) against \( A + B \) and take \( m = 1 / \text{slope} \).

The CECI requires estimates of \( \omega, \phi \) and \( \varepsilon. \) \( \varepsilon \) can be estimated as the population average of \( \varepsilon_{al|ab} \). To this end, we suggest for each dyad of individuals \( a \) and \( b \), where \( a \) is an individual for which we have error free data, we calculate \( M_{a|ab} \) the number of sampling periods in which the standard sampling protocol missed individual \( a \) and we know (from the error free sample) that \( a \) and \( b \) were not together. We can then estimate \( \varepsilon_{a|ab} = \frac{M_{a|ab}}{N_{ab}} \), where \( N_{ab} \) is the number of sampling periods that \( a \) and \( b \) were known not to be together. \( \varepsilon \) can then be estimated as the average of both \( \varepsilon_{a|ab} \) across all dyads containing at least one individual in the error free dataset. Next the researcher can estimate \( \omega \), using the relationship \( \varepsilon_{ab|ab} = (1 + \omega) \varepsilon_{a|ab} \varepsilon_{b|ab} \) in the same manner as suggested for estimating \( \omega \) for the GLECI above.

To estimate \( \phi \) a researcher can use the relationship:

\[
\varepsilon_{a|ab} = \frac{(1 - \phi)(1 + \omega) \varepsilon_{a|ab}^2 \varepsilon_{b|ab}}{\phi(\varepsilon_{a|ab} + \varepsilon_{b|ab})}
\]

For each dyad of individuals for which we have error-free data a researcher can obtain an estimate of \( \varepsilon_{a|ab} = \frac{M_{a|ab}}{N_{ab}} \), where \( M_{a|ab} \) is the number of sampling periods in which the standard sampling protocol missed individual \( a \) and we know (from the error free sample) that \( a \) and \( b \) were together, and where \( N_{ab} \) is the number of sampling periods that \( a \) and \( b \) were known to be together. Using the estimates of \( \varepsilon_{a|ab} \) and \( \varepsilon_{b|ab} \) obtained above, the researcher can then obtain an estimate of \( \varepsilon_{a|ab}^2 \varepsilon_{b|ab} / (\varepsilon_{a|ab} + \varepsilon_{b|ab}) \). \( \varepsilon_{b|ab} \) can be likewise be estimated as \( \frac{M_{b|ab}}{N_{ab}} \), and \( \varepsilon_{b|ab}^2 \varepsilon_{a|ab} / (\varepsilon_{a|ab} + \varepsilon_{b|ab}) \) estimated as for individual \( a \). The researcher can then fit a linear regression (constrained to go through the origin) with the set of \( \varepsilon_{a|ab} \) and \( \varepsilon_{b|ab} \) as the dependent variable and the set of \( \varepsilon_{a|ab}^2 \varepsilon_{b|ab} / (\varepsilon_{a|ab} + \varepsilon_{b|ab}) \) and \( \varepsilon_{b|ab}^2 \varepsilon_{a|ab} / (\varepsilon_{a|ab} + \varepsilon_{b|ab}) \) as the independent
variable. Since the slope of the regression will estimate $(1 - \phi)(1 + \omega)/\phi$, we can estimate $\phi = (1 + \omega)/(\text{slope} + 1 + \omega)$.

A second way we suggest a researcher might obtain estimates of calibration parameters is to have two researchers or research teams independently collecting association data using their standard protocol, for a portion of the data collection period. For example, this might be done by having the second researcher (denoted $Y$) collect data a short time after the first (denoted $X$), on a short enough time scale that group composition is unlikely to have changed. Here we suggest a procedure for obtaining calibration statistics for the CECI since this reduces down to the IIECI or GLECI when the calibration data reveals the relevant component of error to be absent.

We suggest that researchers first obtain estimates of $\varepsilon_{a|ab}$. Ideally we wish to estimate $M_{Xa|ab}/N_{ab}$, the proportion of events that $X$ missed $a$ given $a$ and $b$ were together. We suggest researchers do this by calculating the proportion of sampling periods $X$ missed $a$ given that $Y$ recorded $a$ and $b$ in different groups (which we denote $Y|ab$), i.e. $M_{Xa|Y|ab}/N_{Y|ab}$. We can repeat this procedure, reversing the role of $X$ and $Y$ to obtain $M_{Ya|X|ab}/N_{X|ab}$. We can then take the mean as our estimate, i.e. $\varepsilon_{a|ab} = M_{Xa|Y|ab}/2N_{Y|ab} + M_{Ya|X|ab}/2N_{X|ab}$. $\varepsilon$ can then be estimated as the population average of $\varepsilon_{a|ab}$ as above.

Researchers can then obtain estimates of $\varepsilon_{a|ab}$, the probability that individual $a$ is missed when it is in a group with individual $b$. Again, we have potentially two estimates of this for each combination of $a$ and $b$. First we have $M_{Xa|Yab}/N_{Yab}$, the proportion of sampling periods in which researcher $X$ missed individual $a$ given researcher $Y$ recorded $a$ and $b$ together, and conversely we have $M_{Ya|Xab}/N_{Xab}$. We suggest $\varepsilon_{a|ab}$ be estimated as the average of these, i.e. $\varepsilon_{a|ab} = M_{Xa|Yab}/2N_{Yab} + M_{Ya|Xab}/2N_{Xab}$. We suggest this be done for all combinations of $a$ and $b$ for which $a$ and $b$ were frequently seen together, in order to obtain estimates of $\omega$ and $\phi$. $\omega$ can
be estimated using the relationship $\varepsilon_{ab|\varepsilon_{ab} = (1 + \omega) \varepsilon_{a|\varepsilon_{ab}} \varepsilon_{b|\varepsilon_{ab}}$ in the same manner as suggested for estimating $\omega$ for the GLECI above. $\phi$ can be estimated as above by fitting a linear regression (constrained to go through the origin) with the set of $\varepsilon_{a|\varepsilon_{ab}}$ and $\varepsilon_{b|\varepsilon_{ab}}$ as the dependent variable and the set of $\varepsilon_{a|\varepsilon_{ab}} \varepsilon_{b|\varepsilon_{ab}}/(\varepsilon_{a|\varepsilon_{ab}} + \varepsilon_{b|\varepsilon_{ab}})$ and $\varepsilon_{b|\varepsilon_{ab}} \varepsilon_{a|\varepsilon_{ab}}/(\varepsilon_{a|\varepsilon_{ab}} + \varepsilon_{b|\varepsilon_{ab}})$ as the independent variable yielding the estimate $\phi = (1 + \omega)/(slope + 1 + \omega)$.

DISCUSSION

Studies of animal social networks have shed new light on many ecological and evolutionary processes. For example, the structure of the social environment can shape how information (Aplin, Farine, et al., 2015; Aplin, Farine, Morand-Ferron, & Sheldon, 2012; Farine, Aplin, Sheldon, & Hoppitt, 2015) and diseases (K.L. VanderWaal, Atwill, Isbell, & McCowan, 2013; K.L. VanderWaal et al., in press) spread in wild populations. Further, they have provided important insights into the role of the social environment on shaping selection (Farine & Sheldon, 2015; Formica et al., 2011; McDonald, 2007; Oh & Badyaev, 2010; J.B. Silk, Alberts, & Altmann, 2003; J. B. Silk et al., 2010; Wey, Burger, Ebensperger, & Hayes, 2013).

However, studies have used varying approaches to quantify the relationships among individuals. Whilst care is generally taken to ensure that the chosen approach has biological relevance, the underlying assumptions behind the approach used are almost never explicitly considered. The results of our study into different association indices suggests that many commonly-used approaches should be avoided as they do not accurately estimate the (absolute or relative) strengths of social bonds, which has implications on estimates of social structure and social processes occurring through social networks.

Ideally, studies of animal social networks would capture information about all individuals in the study population at once. Realistically, this is unlikely to be possible in all but a very select number of studies. Thus, before constructing a social
network from a given set of data, we suggest that the following questions should be addressed:

(1) How much data has been collected on each dyad?
(2) Are all individuals sampled equally?
(3) What proportion of the population is observed in each sample?
(4) Are there any mistakes in the observations?

The issues surrounding question 1 have now been relatively well outlined in the literature (Farine & Strandburg-Peshkin, 2015; Franks et al., 2010; Lusseau, Whitehead, & Gero, 2008; M. J. Silk et al., 2015; Whitehead, 2008). In general, these studies have found that collecting enough data on each dyad (at least 20 observations per dyad) is important for accurately estimating global social network structure. In the current study, we address issues arising from questions 2 and 3, and how better indices can reduce the potential impact that missing observations can have on both the absolute and relative estimates of association strengths among individuals. Question 4 represents an area requiring some further investigation.

Our study suggests that a critical step in the study of animal social networks will be the collection of calibration data. Currently-used association indices are all based on arbitrary rates of missing observations. For example, the half-weight index assumes that the probability of missing individuals a and b when they are together is exactly half the probability of missing either individual when they are apart. Importantly, we have shown that when this is not true, the HWI does not result in a ‘better approximation’ of the real association rate when compared to the simple ratio index. Thus, we recommend avoiding the use of the HWI, and instead using the SRI when no calibration data is available (see Figure 3). In reality, it is likely that the rates of observation could be estimated from parameters of the observation data, such as the average group size, the average number of individuals observed in a sampling period, and the average number of groups sampled. Whether these can be used to parameterize the M-weighted index warrants further investigation.
Several extensions of association indices have been proposed to deal with other issues arising when sampling populations. Godde, Humbert, Cote, Reale, and Whitehead (2013) suggest a method to correct for the fact that individuals that prefer large groups are more likely to be observed together. This involves normalizing the association index values by the two individual’s combined gregariousness (the sum of their association indices to others). To deal with other potentially confounding influences on association patterns (such as home-range overlap) when attempting to estimate true association rates, Whitehead and James (2015) propose regressing association indices against other input parameters. Our proposed indices work equally well with both of these approaches as they are simply new ways of defining the association value for pairs of individuals. The above two studies highlight how patterns of affiliation can be affected by a range of different factors. Thus, even if good calibration data can be obtained to estimate accurate relationship strengths, it will always be important to use null models when conduction hypothesis testing with animal social networks (Farine in review).

We encourage further investigation into methods for collecting informative calibration data alongside the social network data. One potential avenue could be to use mark-recapture techniques that explicitly investigate detection probabilities, and these could be conditioned on having observed one or more particular individuals. There are also increasing numbers of studies that are collecting complete datasets from groups or populations of animals, and these could provide very useful data for testing different approaches to collect calibration data. A particular challenge that will arise is that social network analysis has proved particularly useful in species or communities that exhibit fission-fusion dynamics (Aureli et al., 2008; Couzin, 2006; M. J. Silk, Croft, Tregenza, & Bearhop, 2014). Here, the rate of turnover in group membership can be very rapid (for example group membership in great tits, Parus major, can be close to random after just 10 minutes, Farine, Firth, et al., 2015). Similarly, when using focal observations—following a single individuals and recording its interactions with others—the concept of a
‘group’ is unclear, and we do not yet have a definition for out to estimate the group location error. This is because all individuals can be observed, but only interactions among a subset of edges (those connected to the focal) are recorded. Thus, in such systems, more research is required to find robust approaches for collecting calibration data.

In all of our simulations, and more generally in the assumptions of how well any analysis captures reality, our estimates of accuracy are a ‘best-case scenario’. In reality, most datasets will also contain erroneous observations. At best, these are simply individuals wrongly assigned into a group in which they did not occur. If such individuals (and their erroneous associates) are observed many times, then the resulting association strength will be low, and the error will be reasonably well dealt with by using weighted social networks. However, the impact of incorrectly assigning one individual identity for another could be significantly greater, especially as such errors are unlikely to be randomly distributed throughout the dataset (i.e. certain pairs of individuals are more likely to be confounded than others). No association index will be able to correct for such errors. However, the relative effect that incorrect assignments of identity have on different approaches to estimate social network structure warrants further investigation.

Our paper makes it clear that we should avoid blindly using association indices without proper consideration of the assumptions that they entail. We also recommend discontinuing the use of the HWI, which Cairns and Schwager (1987) already 30 years ago identified as being problematic because of the many assumptions it makes. Instead, we recommend using properly calibrated association indices wherever possible, and using the SRI if no appropriate calibration data or estimates of rates of detectability are available. Whichever approach is used, we hope that our paper will at least encourage researchers to carefully and explicitly consider their choice of approach for estimating association strengths among individuals in their study population.
ACKNOWLEDGEMENTS

DRF received additional support from the BBSRC (BB/L006081/1 to B.C. Sheldon).

REFERENCES


