Social Network Analysis and Valid Markov Chain Monte Carlo Tests of Null Models

Krause S*, Mattner L[α], James R[†], Guttridge T[&], Corcoran MJ, Gruber SH** & Krause J[&]**

*Fachbereich Elektrotechnik und Informatik, University of Applied Sciences Lübeck, 23562 Lübeck, Germany
[α]Institut für Mathematik, Universität zu Lübeck, 23560 Lübeck, Germany
**Bimini Biological Field Station, 15 Elizabeth Drive, South Bimini, Bahamas
[†]Department of Physics, University of Bath, Bath BA2 7AY, UK
[&]School of Biology, University of Leeds, Leeds LS2 9JT, UK.

5th February 2008
Revised 26th August 2008

Corresponding Author: Stefan Krause, Fachbereich Elektrotechnik und Informatik, University of Applied Sciences Lübeck, 23562 Lübeck, Germany.
Email: krause@fh-luebeck.de
Phone: +49 451 300 5315
Fax: +49 451 300 5236

Keywords

Null models, social network analysis, Markov chain Monte Carlo tests, group living.

Abstract

Analyses of animal social networks derived from group-based associations often rely on randomisation methods developed in ecology (Manly 1995) and made available to the animal behaviour community through implementation of a pair-wise swapping algorithm by Bejder et al. (1998). We report a correctable flaw in this method, and point the reader to a wider literature on the subject of null models in the ecology literature. We illustrate the importance of correcting the method using a toy network, and use it to make a preliminary analysis of a network of associations among eagle rays.

Introduction

The field of animal behaviour is increasingly turning to a network approach to study social structure on all levels from the individual to the population (Croft et al. 2008). Many of the animal social networks being reported are for group-living species, and are constructed using what Whitehead and Dufault (1999) dubbed the "gambit of the group". All animals in an observed group are assumed to be associating, then these associations are accumulated over a chosen observation period to yield the network.

A key feature of this approach is that, even if the membership of groups within a population is entirely "random", still the accumulated network could contain apparent "non-random" structure. Then in order to deduce virtually anything of biological interest from the network it is necessary to compare the observed structure against some form of null model of network formation (Croft et al. 2008). Here we report on some problems and statistical issues associated with some of the null models that are being used in animal social network studies, and what we can learn from ecologists and statisticians about how to solve them. We also comment on whether more thought should sometimes be given to the construction of the null model itself, and illustrate our points with a simple toy network and a real animal social network of a group-living species.

Null models in network theory typically consist of randomizations or so-called Monte Carlo simulations. An empirical data set is compared to itself in series of randomised versions which forms the starting point of network data analysis in most cases (see Manly 2007 for details of randomisation procedures). Let's say we use the descriptive statistic Z which measures a certain property of networks that we are interested in. The comparison of the value of Z for the empirically measured network compared to the values found in randomised versions of this network tells us whether we should attach particular importance to this test statistic; i.e. whether or not it is significant at the 5% level. At first glance it may seem that there is only one way in which to randomise a network. However, this is not the case because we are usually interested in the question of whether a particular aspect of a biological system (and not the entire system itself) could have arisen by chance or whether it is the product of a biological mechanism which in a next step we then hope to identify. The art of developing null models lies in the decision of which components of a biological system to keep

constant and which ones to randomise. At this point it should be clear that proper conclusions from social networks can only be drawn if subjected to an appropriate randomisation test.

Main Issues in the Analysis of Social Networks

Let's start by looking at how data are typically collected by investigators. In some biological systems it is possible to survey an entire area (a whole pond or an entire hill-side) to obtain information on the association patterns of all individuals in the population at a given moment in time (examples of this include red deer Croft et al. 2008; guppies Croft et al. 2006; sea-lions Wolf et al. 2007). This is usually the case with human social networks as well. We will leave those cases aside because their analysis is well understood and can be performed as described by Croft et al. 2008.

However, not all animals are this easy to observe and in particular many large animal species such as most marine mammals, non-human primates and ungulates are often encountered only in the form of a single group (or a few groups) per day (or other time unit) which means we have no information on what the rest of the population is doing at this point in time. Records of different groups thus come from different sampling dates/times. A set of such records forms the network sample that is subject to a statistical analysis. A common approach to investigating the social structure of such species is to perform random swaps of pairs of individuals from different groups which if carried out over many steps should result in some sort of randomisation of the data set.

We believe there are four issues worthy of attention:

1. The null model, i.e. a set of possible network samples (including the observation) which are regarded equally probable, if the biological mechanism we are interested in is not present.
2. The descriptive statistic that measures the effect of the biological mechanism by assigning scores to network samples.
3. The mathematical test that determines whether the statistic value of the observation significantly differs from random.
4. The algorithm used to generate randomised network samples (i.e. members of the null model) with equal probability.

Null models

Quite generally, a null model may be defined by choosing certain features of systems under study, and declaring all systems with the same values of these features as equally likely. In the specific case we are concerned with, group-based association data can be, and often is, represented using a presence-absence matrix. Each row (say) represents an observed group and each column an animal. Entries in the matrix are 1 if the animal is seen in that group, and 0 otherwise. A common null model for such data is defined by declaring as equally likely all presence-absence matrices with the same row and column sums. That is, the group sizes and the number of times each animal is observed are both constrained, but who belongs to which group is scrambled. This null model should then in principle be tested by comparing some suitable test statistic evaluated for the observed matrix with the distribution of the test statistic under the uniform distribution on all matrices with the same row and column

sums as the observed one. This approach, modified by a Monte-Carlo method necessary due to the computational intractability of the test distribution, was introduced into ecology by Manly (1995), where the entries in the matrix indicated, for example, the presence or absence of species on islands.

The work of Manly (1995) was the primary source for methods developed for the analysis of animal social networks (Bejder et al. 1998), where the entries in the matrix denote the presence or absence of individuals in groups. However, since that time there has been quite a debate in the ecology literature (Gotelli & Entsminger 2001; Manly & Sanderson 2002; Gotelli & Entsminger 2003; Miklós & Podani 2004; Lehsten & Harmand 2006) about the rights and wrongs of various methods purporting to generate random presence-absence matrices. This debate appears not to have been picked up in the animal network and association literature, despite the fact that the early methods formed the basis for software programs that have become available for network analysis. The aim of this article is first of all to inform the animal behaviour community about the existence of such a debate and its crucial points. This information has important implications for the correct use of network theory for the analysis of the social organisation of animals, in particular if the common null model described above is used. Therefore, this article mainly deals with approaches based on this particular and frequently used null model. Our second aim is to draw attention to the fact that even if correctly applied some of the methods advocated in the ecological literature may not necessarily be the most appropriate when it comes to analysing social networks. The underlying biological mechanisms, the observation method, and other factors may have an influence that should be taken into account when defining null models. In this article we will familiarise the reader with some of the more important points of null model development and illustrate potential dangers of incorrect applications with example data sets.

Test statistics

Given any null model and any data to challenge it, a test should ideally be chosen so as to have maximal power, that is, rejection probability, for scientifically relevant alternatives. Unfortunately, the latter are hard to specify for complex data structures like our presence-absence matrices. Even if the relevant alternatives were specified, there would not usually exist a test having maximal power for each of these, and it would then be a challenging problem for a mathematical statistician to devise a test having at least reasonable power under each. It is common in many fields to ignore this problem or, speaking more charitably, to "solve" it pragmatically by proposing some test statistic on intuitive grounds. Specifically for presence-absence matrix data, an example used frequently in ecology is the C-score (see for example Lehsten & Harmand 2006), which we adopt here for illustrative purposes.

The C-score measures the mean number of so-called checkerboard units in a presence-absence matrix. A checkerboard unit is a 2 x 2 submatrix with ones in one pair of opposite corners and zeros in the other. For a matrix M with k species (or individuals) the C-score is the number of checkerboard units in M divided by the number $k(k-1)/2$ of unordered pairs of species (or individuals).

Now, choosing the C-score or any other test statistics implicitly defines the relevant alternatives to the null model as those for which the resulting test has at least

moderate power, say 0.8 or more. It is then again a challenging mathematical task to describe these alternatives in a way that enables the judgement of their real scientific relevance. This problem will not be addressed in this paper.

Test methods

The standard Monte Carlo (SMC) approach

From now on we suppose that we wish to test the common null model for presence-absence matrices with some specified test statistic T, rejecting for large values of $T(X_1)$ with X_1 being the observed matrix. The distribution of T under the null model being computationally intractable for all but the smallest matrices, it seems natural to apply the SMC approach, due to Dwass (1957) and Barnard (1963), which in our case would require simulating a number n-1 of matrices X_2, ..., X_n independently and uniformly distributed over the set $\mathcal{X}$ of all matrices with the same row and column sums as the observed one, and to reject or retain the null hypothesis according to the rank of $T(X_1)$ within $T(X_1)$, ..., $T(X_n)$. More precisely, given a level α, we would reject the null hypothesis if $T(X_1)$ were among the $n*\alpha$ largest of the $T(X_i)$. It is known from general results that such an SMC test keeps the given level α exactly under the null hypothesis and, if n is sufficiently large, approaches the power of the original non-Monte-Carlo test, see Jöckel (1986).

There are algorithms that construct matrices X_i as above, each from scratch by filling an empty matrix, using backtracking to meet the randomisation constraints. These algorithms are difficult to implement efficiently. Care has to be taken in order to make sure that all matrices are generated with equal probability. Gotelli & Entsminger (2001) describe what mistakes have been made in the past, and suggest a correction to a fill algorithm, called "Knight's Tour".

The quasi Monte Carlo (QMC) approach

The following approach has been suggested to bypass the difficulties involved in performing a SMC test: Starting with the observation X_1, we simulate a Markov chain $X_1, X_2, \ldots, X_n$ on the set $\mathcal{X}$, using some available transition probabilities under which the uniform distribution on $\mathcal{X}$ is invariant. This holds for the transition probabilities of an algorithm suggested by Besag & Clifford (1989), described below. Then, under the hypothesis of the uniform distribution of X_1, the X_i will be uniformly distributed but unfortunately dependent.

Hence, there is no guarantee that the SMC test described in the previous section is applicable. Therefore, in order to diminish the dependence between generated matrices, intermediate steps may be applied, such that only every kth matrix in a chain (with k-1 intermediate steps) will be put into the subset that is used for the ranking. Lehsten & Harmand (2006) apply a procedure to estimate a suitable number of intermediate steps developed by Raftery & Lewis (1996). Another heuristic of this kind was described by Miklós & Podani (2004). However, it is unknown whether any of these numbers of intermediate steps suffice to make the SMC test applicable.

A valid Markov chain Monte Carlo (MCMC) approach

As an alternative to QMC testing, Besag and Clifford (1989) proposed two methods, called the parallel and the serial methods, that were specifically designed for Markov chain Monte Carlo testing. Manly (1995) uses the serial method (however, with a biased Markov chain construction as will be explained in the next section). Miklós & Podani (2004), and Lehsten & Harmand (2006), who point out the relevance of a correct Markov chain construction, do not mention this test method, but use the QMC test instead.

Since the parallel and serial methods provide exact tests, they are certainly worth being considered. Let us briefly describe here the latter. For the serial method, we choose an integer m randomly and uniformly from the numbers 1,...,n and define X_m to be the actually observed matrix. Using the available transition probabilities n-m times, we construct a Markov chain X_m, ... , X_n, and using the corresponding reverse transition probabilities m-1 times, we construct a Markov chain X_m, ... , X_1. We then reject or retain the null hypothesis according to the rank of $T(X_m)$ within $T(X_1)$, ..., $T(X_n)$, analogously to the standard Monte Carlo test. This is a valid test, since under the null hypothesis the random index m happens to be independent of the sequence $X_1,...,X_n$. We refer to Kemeny & Snell (1983) for the elementary Markov chain theory just assumed.

Constructing a correct Markov chain

The methods discussed above depend on constructing a correct Markov chain. Besag & Clifford (1989) suggested a swapping algorithm that can be used to achieve this. However, a number of papers, including Manly (1995) and Bejder et al. (1998), contain a description of this algorithm that is not correct and leads to biased randomisations (Miklós & Podani 2004). It is worth looking at the problem more closely.

At first, we give a description of the correct swapping algorithm that follows Besag & Clifford (1989). For a matrix X_1 randomly choose a 2 x 2 submatrix. If this submatrix forms a checkerboard as defined above, exchange its ones and zeros, and let X_2 denote the resulting matrix. If not, let $X_2 = X_1$. Proceeding this way, a Markov chain is constructed. Steps in this chain are sometimes called “trial-swaps” and more specifically "waiting steps" if $X_2 = X_1$, and "successful swaps" otherwise. Besag & Clifford (1989) showed that this chain has as its invariant distribution the required uniform distribution on $\mathcal{X}$, with Manly (1995) observing that reference should have been made to a nontrivial auxiliary result of Ryser (1957) apparently taken for granted by Besag and Clifford.

Manly (1995) and Bejder et al. (1998), however, construct a Markov chain by randomly selecting checkerboard units instead of arbitrary 2 x 2 submatrices. This change eliminates waiting steps from the Markov chain described by Besag & Clifford (1989). As a consequence, equidistribution of the generated matrices cannot be guaranteed.

Miklós & Podani (2004) give an illustrative example, consisting of five 3x3 matrices, that shows the differences in the transition probabilities between the algorithm

described by Manly (1995) and the algorithm of Besag & Clifford (1989) (though without reference to the latter).

It is hard to say what the practical consequences of the biased swapping algorithm are, because this depends on the structure of the network sample, and of the test statistic that is used. Lehsten & Harmand (2006) analysed 291 published presence-absence matrices. They report that only in 5 cases would the biased swapping have led to a different statistical decision compared to the correct swapping algorithm. The test statistic they used for their evaluation was the ecologically relevant C-score (which we described in the section on test statistics).

To illustrate possible consequences regardless of a specific test statistic, we analysed an artificial example with 13 individuals distributed to 4 groups of sizes 2, 2, 2, and 8. One individual occurs twice, i.e. in two groups, all others occur once. Given these settings, there are 89100 matrices with the same row and column sums, which can be divided into two types. Type A contains the individual that occurs twice in small groups (size 2) only, in matrices of type B this individual occurs in the large group (size 8). Of the matrices 20% (17820) are of type A, and 80% (71280) of type B. The biased swapping algorithm, however, will generate 21.8% of type A and 78.2% of type B.

We observed this tendency towards over-sampling matrices that contain frequently occurring individuals in the small groups in several other experiments we have run.

If a test statistic like the C-score is used, which in our example assigns the value 0.7308 to matrices of type A, and the value 0.6538 to matrices of type B[1], a test using the biased swapping algorithm will be slightly conservative, because matrices of type A will be generated more often than they should. For other test statistics the test results may be biased in a different way.

A small simulation study

To illustrate the differences between the QMC and the MCMC approach we ran both the QMC test and the serial MCMC test on a number of randomly chosen matrices from the set of 89100 matrices described above. Instead of the C-score, which only yields two different values for this set, we used an artificial score[2] that produces about 14700 different values. This was done to obtain a closer relationship between matrices and scores and to avoid too many ties during the ranking. If ties occur, these have to be broken randomly. With only two different score values the rank positions would

[1] Matrices of type A have 57 checkerboards, and matrices of type B have 51 checkerboards. Since there are $13 * 12/6 = 78$ unordered pairs of individuals in our example, the C-score of type A matrices is $57/78 = 0.7308$, and the C-score of type B matrices is 0.6538.

[2] Each individual is assigned a prime number. For each group, i.e. each row in the matrix, a score is computed by multiplying the prime numbers of all individuals in this group. The score of a matrix is the sum of all row scores.

strongly be influenced by this procedure and the effects we are demonstrating would be less obvious. Although this score certainly does not measure a biologically relevant feature, it serves as an appropriate substitute in a simulation based on a very small network, where biologically meaningful scores would yield only a few different values.

In order to get comparable results we chose the parameters for each test such that both generated roughly the same numbers of matrices. Also, since the network represented by the matrices is rather small (4 groups, 13 individuals) we started with small numbers of generated matrices (100, which is about 0.001% of 89100). For real networks the percentage of possible matrices covered by the test will be much smaller. Lehsten & Harmand (2006) refer to an actually observed 4 x 180 matrix, for which $4.7 * 10^{68}$ matrices with the same row and column sums exist.

We set the level α to 5% and ran each test on 10^7 randomly drawn "observations" from our set of 89100 matrices. A test that keeps its level would reject the null hypothesis in about (5±0.007)% of the cases, where $0.007 = 100 * \sqrt{0.05 * 0.95/10^7}$ is the standard deviation in percent.

Table 1 shows the results for different parameter settings.

N denotes the number of matrices (including the initially drawn matrix, which in practise would be the observed one) that are used for the ranking. *IS* denotes the number of steps ("trial-swaps") between each two of the *N* matrices. The product *N***IS* specifies the total number of generated matrices. If *IS* = 1, then for all matrices a score will be computed and included in the ranking. For *IS* = k, k > 1, only every kth matrix will be used.

The results in Table 1 show that the serial MCMC method keeps its level exactly regardless of the number of generated matrices, in accordance with what Besag & Clifford (1989) proved. The QMC test (with the correct swapping algorithm) begins to work merely approximately only when sufficiently many intermediate steps are introduced.

It is also important to look at the power of the tests, i.e. the probability to correctly reject the null hypothesis. Again, we ran tests on 10^7 randomly drawn matrices based on the same network type as before, but this time we modified the random drawings such that the 5% matrices with highest scores were drawn with a probability of 80%.

Table 2 shows the results for the same parameter settings as in Table 1. The power of both test methods increases as the number of generated matrices increases. The relatively high values for the QMC test in the first two lines are most probably mainly due to the fact that with these parameter settings the test does not keep its level.

The values in the last line of in Table 2 seem to suggest that the MCMC test has higher power than the QMC test, when a total number of 10,000 matrices are computed (including all intermediate matrices). This is not entirely true. It turns out that with *N* = 1000 and *IS* = 10 we get 72.48% rejections from the QMC test, which with these settings also roughly keeps its level (5.07%). However, in practice it is

difficult to decide which parameter settings for *N* and *IS* should be chosen, if only limited time for running the test is available. Miklós & Podani (2004) suggest that the number of intermediate trial-swaps should be set such that the expected number of actual swaps equals twice the number of ones in the matrix. Of course, the same number of intermediate steps has to be used between any two matrices used in the ranking. Therefore, this number has to be computed from one matrix, for example the observed one, and then be used for all sequences of intermediate steps. This heuristic in our example suggests setting *IS* to 230 or 257, depending on whether we start with a matrix of type A or type B. These settings only slightly increase the power compared to doing 100 intermediate steps. For *IS* = 100 we get 66.41%, and for *IS* = 230 we get 66.52%.

Application to some real data

In the following we will provide a case study that illustrates the MCMC-approach using a data set on spotted eagle rays (*Aetobatus narinari*). The data were collected in the costal waters around Bimini, Bahamas, between January and November of 1999. Bimini comprises two small islands which enclose a 16 km^2 mangrove-fringed lagoon (Figure 1). The Bimini Islands are frequented by approximately 200 eagle rays which often form large schools of up to dozens of individuals. The spot pattern on the back of the rays allows individual photo-id (see details of method in Corcoran & Gruber 1999). Different sampling sites around the islands were visited on at least a weekly basis and surveyed for eagle rays. Following Pitcher & Parrish (1993) we defined a "social group" as two or more rays swimming in an apparently coordinated manner within approximately three to four body lengths of one another.

We do not analyse the data in full detail here, but rather use one specific test statistic as an example. It is based on Association Strength (AS) (Croft et al. 2008), which measures the number of times a pair of individuals is seen in the same group in the data set. In order to turn this into a simple test statistic that does not introduce new parameters we computed the sum of squares of AS values for all possible pairs in a matrix. This AS-square-sum score emphasizes large AS values (in contrast to, for example, a linear sum of AS values without weights, which would yield the same score value for all randomised matrices).

The original data set contains 54 groups with 159 different rays sampled at 13 sites. The sum of all group sizes is 372.

For our example analysis we restricted the data set to samples from two main sites shown in Figure 1, which have a short distance from each other. This was done because the eagle rays show a tendency towards site fidelity. (Further analysis will be necessary to check whether this restriction provides a basis for meaningful conclusions from test results.) The restricted data set contains 26 groups with 81 different rays sampled on 20 different days. The sum of all group sizes is 154, and the mean group size is 5.92. The largest group contains 22 rays, the smallest groups 2 rays.

This means that the observation is represented by a 26×81 matrix with 154 ones. It has 8662 checkerboards and 1053000 2×2 submatrices. The observation matrix contains 593 pairs occurring in one group only, 54 pairs occurring in two groups, and

8 pairs occurring in three groups. The AS-square-sum score is $1*593 + 4*62 + 9*8 = 881$.

The results of repeated serial MCMC tests with different numbers N of generated matrices are shown in Table 3. As can clearly be seen, the results are not significant at the 0.05 level. The differences between p-values of multiple tests for the same N decrease as N increases. The running time grows linearly with the size of N. Even for large values of N the running times look moderate.

The power of the test should increase, if more matrices are generated, or if some intermediate trial swaps are performed between each two ranked matrices. It has further to be explored which numbers should be chosen in order to make the best use of the available time for running the test. In contrast to the QMC test, however, intermediate steps are not required to make the test "more exact". This means, there is no risk of performing "not enough" of them.

A short note on the software and hardware used to run the tests:
The test software was written in a straightforward way without optimisations, except for a few trivial ones. Optimisations of the code would significantly reduce the running times. The programming language Java was used (JDK™ and Java™ SE Runtime Environment 1.6.0_03 from Sun Microsystems), and the software was run on Windows XP® on a modern desktop computer.

Discussion

Given the common null model (presence-absence matrices with the same row and column sums), we strongly suggest using the serial Markov chain Monte Carlo test of Besag & Clifford (1989) as already proposed by Manly (1995). This is always a valid test regardless of the number of matrices generated, and it has shown good power in our small simulation study (with a 3×8 matrix) depending on a given matrix and some test statistic. Whether it has similar power properties for other test statistics and other matrix types remains to be investigated. Also, it might be worth taking into consideration the parallel test of Besag & Clifford (1989), and comparing it to the serial one, which we have not done in this paper. Simulation studies on a lager scale may help to give more concrete advice on which test should be chosen in which situation.

We have not found any advantages of the commonly applied quasi Monte Carlo test (as used for example by Bejder et al. 1998) over the serial Markov chain Monte Carlo test. The quasi Monte Carlo test is at best approximately valid, even if the swapping is done correctly. Also, in practice it can be difficult to choose the numbers of generated matrices and of intermediate steps properly, because these can depend on the structure of the sampled network, and on the test statistic chosen.

Also, regardless of the test used care has to be taken to choose a correct construction algorithm for the Markov chain. The algorithm described for example by Manly (1995) and Bejder et al. (1998) is biased and does not provide a basis for any of the Monte Carlo tests that make use of Markov chains without further modifications. In practice, the bias leads to over-sampling matrices where frequently observed individuals occur in small groups. A correct algorithm, as explained in this paper, was

first described by Besag & Clifford (1989) and rediscovered by Miklós & Podani (2004).

The analysis of social networks is still a challenging task consisting of different sub-problems. It is still unknown whether the commonly used null model is always appropriate, when applied to randomly observed groups. The underlying biological mechanisms, the observation method, and other factors may have an influence that should be taken into account. Also, it should be noted that the use cases in ecology and social network analysis are different. At each point in time one species may be present on several different islands. With occurrences of individuals in groups this is different. It is difficult to say to what degree the biological mechanisms that lead to the observed networks are relevant for the development of null models. However, it should not be taken for granted that a good null model for one case is also an appropriate null model for another case. This not only regards sampling dates and times, which may impose restrictions on what forms a meaningful sample network, but also the constraints that define the elements of the null model. Lehsten & Harmand (2006) suggest that in spite of null models with constant row and column sums having frequently been used the design of null models should take into account "the ecological mechanism of interest and the available data."

Acknowledgements

JK acknowledges financial support from the NERC and the EPSRC and TG was supported by a Leverhulme fellowship. SK is grateful to Ralf Schiffer for some very useful initial discussions. We thank Herbert Krause for providing the drawing in Figure 1.

References

Barnard, G. A. (1963). Discussion of a paper by M. S. Bartlett. *Journal of the Royal Statistical Society, Series B* **25**: 294.

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

Besag, J. E. and P. Clifford (1989). Generalized Monte Carlo significance tests. *Biometrika* **76**: 633-642.

Corcoran M.J. and Gruber S.H. 1999. The use of photoidentification to study social organization of the spotted eagle ray, *Aetobatus narinari* (Euphrasen 1790), at Bimini, Bahamas: a preliminary report. Bahamas J. Sci. **7**(1):21-27.

Croft, D.P., R. James and J. Krause (2008). *Exploring Animal Social Networks*. Princeton University Press, Princeton.

Croft, D.P., R. James, P. O. R. Thomas, C. Hathaway, D. Mawdsley, K. N. Laland and J. Krause (2006). Social structure and co-operative interactions in a wild population of guppies (poecilia reticulata). *Behavioral Ecology and Sociobiology* **59**(5): 644-650.

Dwass, M. (1957). Modified randomization tests for nonparametric hypotheses. *Annals of Mathematical Statistics* **28**: 181-187.

Gotelli, N. J. and G. L. Entsminger (2001). Swap and fill algorithms in null model analysis: rethinking the knights tour. *Oecologia* **129**: 281-291.

Jöckel, K.-H. (1986). Finite sample properties and asymptotic efficiency of Monte Carlo tests. *Annals of Statistics* **14**: 336-347.

Kemeny, J.G. and J.L. Snell (1983). *Finite Markov Chains*. Springer, New York.

Lehsten, V. and P. Harmand (2006). Null models for species co-occurrence patterns: assessing bias and minimum iteration number for the sequential swap. *Ecography* **29**: 786-792.

Manly, B. F. J. (1995). A note on the analysis of species co-occurrences. *Ecology*, **76**: 1109-1115.

Manly, B. F. J. (2007). *Randomization, Bootstrap and Monte Carlo Methods in Biology*. 3rd ed., Chapman \& Hall/CRC, Boca Raton.

Manly, B. and J. G. Sanderson (2002). A note on null models: justifying the methodology. *Ecology* **83**: 580-582.

Miklós, I. and J. Podani (2004). Randomizations of presence-absence matrices: comments and new algorithms. *Ecology* **85**: 86-92.

Pitcher T.J. and Parrish J.K. 1993. Functions of shoaling behavior in teleosts. pp. 363–439. *In*: Pitcher TJ (ed.) Behaviour of Teleost Fishes, Chapman & Hall, London.

Raftery, A. E. and S. M. Lewis (1996). Implementing MCMC. In: Gilks, W.R. et al. (eds), *Markov chain Monte Carlo in Practice*. Chapman and Hall, pp. 115-130.

Ryser, H. J. (1957). Combinatorial properties of matrices of zeros and ones. *Canadian Journal of Mathematics* **9**: 371-377.

Wolf, J.B.W., D. Mawdsley, F. Trillmich and R. James (2007). Social structure in a colonial mammal: Unravelling hidden structural layers and their foundations by network analysis. *Animal Behaviour* **74**: 1293-1302.

QMC			serial MCMC		
N	*IS*	rejected	*N*	*IS*	rejected
100	1	10.195%	100	1	5.007%
100	5	6.509%	500	1	4.995%
100	10	5.691%	1000	1	4.990%
100	100	5.016%	10000	1	5.004%

Table 1. Results from running the QMC test and the serial MCMC test at level 5% on 10^7 randomly drawn matrices from a null model with 89100 matrices.

QMC			serial MCMC		
N	*IS*	rejected	*N*	*IS*	rejected
100	1	39.64%	100	1	22.03%
100	5	55.36%	500	1	46.30%
100	10	59.81%	1000	1	55.44%
100	100	66.41%	10000	1	72.06%

Table 2. Results from running the QMC test and the serial MCMC test at level 5% on 10^7 randomly drawn matrices from a collection that contains 423000 matrices, 80% of which have a score that is in the 5% highest scores.

N	$5*10^7$	10^8	$2*10^8$	$5*10^8$	10^9	$5*10^7$ (IS=10)
p-value 1	0.0717	0.0740	0.0695	0.0739	0.0715	0.0714
p-value 2	0.0718	0.0729	0.0725	0.0724	0.0719	0.0723
p-value 3	0.0743	0.0705	0.0704	0.0714	0.0711	0.0728
p-value 4	0.0756	0.0724	0.0692	0.0726	0.0719	0.0719
p-value 5	0.0662	0.0717	0.0702	0.0720	0.0734	0.0720
p-value 6	0.0763	0.0744	0.0697	0.0705	0.0715	0.0720
p-value 7	0.0703	0.0727	0.0733	0.0720	0.0725	0.0732
p-value 8	0.0733	0.0740	0.0733	0.0725	0.0719	0.0725
p-value 9	0.0674	0.0743	0.0721	0.0722	0.0718	0.0714
p-value 10	0.0696	0.0712	0.0730	0.0728	0.0726	0.0718
Mean p-value	**0.0716**	**0.0728**	**0.0713**	**0.0722**	**0.0720**	**0.0721**
Mean Running time	28 s	56 s	112 s	281 s	562 s	271 s

Table 3. Results from applying the serial MCMC test to the eagle ray data. Each test was run 10 times for different numbers N of generated matrices. The table contains the p-values, the mean p-value, and the running time for a single test (which was almost the same for each of the 10 repetitions). The last column shows results for tests where additional intermediate trial swaps were performed.

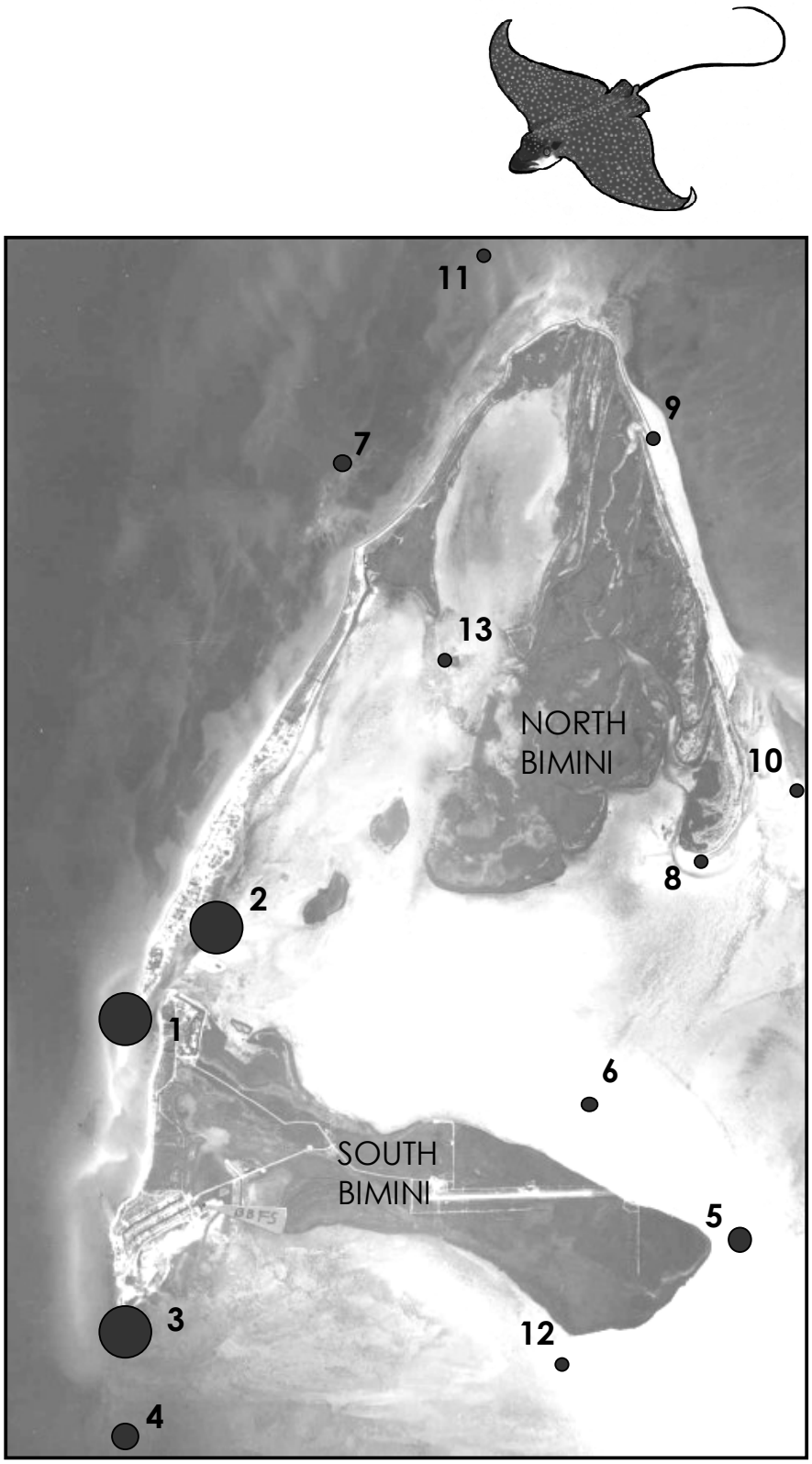

Figure 1. The Bimini Islands. The sampling sites of eagle rays are marked with circles, where the size of each circle corresponds to the number of groups observed at this site. The locations 1 and 2 were chosen for our case study because they contained a large percentage of all individuals (51%) and of all groups (48%).