



# Bayesian modelling and computation utilising directed cycles in multiple network data

Anastasia Mantziou<sup>1</sup> · Sally A. Keith<sup>2</sup> · David M.P. Jacoby<sup>2</sup> · Simón Lunagómez<sup>3</sup> · Robin Mitra<sup>4</sup>

Received: 7 July 2025 / Accepted: 18 December 2025  
© The Author(s) 2026

## Abstract

Modelling multiple network data is crucial for addressing a wide range of applied research questions. However, there are many challenges, both theoretical and computational, to address. Network cycles are often of particular interest in many applications; for example in ecology a largely unexplored area has been how to incorporate network cycles within the inferential framework in an explicit way. The recently developed Spherical Network Family of models (SNF) offers a flexible formulation for modelling multiple network data that permits any type of metric. This has opened up the possibility to formulate network models that focus on network properties hitherto not possible or practical to consider. In this article we propose a novel network distance metric that measures similarities between networks with respect to their cycles, and incorporates this within the SNF model to allow inferences that explicitly capture information on cycles. These network motifs are of particular interest in ecological studies aimed at understanding competitive and hierarchical interactions. We further propose a novel computational framework to allow posterior inferences from the intractable SNF model for moderate-sized networks. Lastly, we apply the resulting methodology to a set of ecological network data studying aggressive interactions between species of fish. We show our model is able to make cogent inferences concerning the cycle behaviour amongst the species, and beyond those possible from a model that does not consider this network motif.

**Keywords** Doubly intractable distributions · Importance Sampling · Object data analysis · Relational data

## 1 Introduction

In many fields, modelling network data is essential to answering the applied research questions of interest. In ecology, the interactive behaviour of different individuals or species within a geographical area can be represented by a directed network with each species corresponding to a node and the edges representing interactions between species (Delmas et al. 2019; Mittelbach and McGill 2019). If these interactions are directed, such as aggression behaviour between individuals of different species, these networks can capture competitive species interactions at the ecosystem scale, e.g. certain species vying over a particular food source or type of habitat. Each aggressive interaction can be represented through a directed edge (i.e. arrow) with the direction indicating which species is the aggressor and which is the recipient respectively.

Within ecology, a particular phenomenon of interest is where the aggressive interactions result in intransitive competition patterns, a set of cyclical interactions that results in no single dominant species, with different species winning

✉ Anastasia Mantziou  
anastasia.mantziou@warwick.ac.uk

Sally A. Keith  
sally.keith@lancaster.ac.uk

David M.P. Jacoby  
d.jacoby@lancaster.ac.uk

Simón Lunagómez  
simon.lunagomez@itam.mx

Robin Mitra  
robin.mitra@ucl.ac.uk

<sup>1</sup> Department of Statistics, University of Warwick, Coventry CV4 7AL, U.K.

<sup>2</sup> Lancaster Environment Centre, Lancaster University, Lancaster LA1 4YQ, U.K.

<sup>3</sup> Departamento Académico de Estadística, Instituto Tecnológico Autónomo de México (ITAM), Río Hondo, Ciudad de México 01080, Mexico

<sup>4</sup> Department of Statistical Science, University College London, Gower Street, London WC1E 6BT, U.K.

out depending on the circumstances. Intransitive competition is of ecological importance as it is thought to promote species coexistence (Laird and Schamp 2006). In a network these would be characterised through directed cycles (Koutrouli et al. 2020).

More generally, cycles reveal information about network topology (Maugis et al. 2017; Fan et al. 2019), and is a motif of interest in many applications beyond ecology. Examples include neuroscience, where the formation of cycles in a human brain network is crucial for human cognitive functions (Sizemore et al. 2018), and biology where RNAs forming covalently closed loop structures, called circular RNAs, have been associated with diseases such as cancer (Han et al. 2017).

Recent methodological developments permit data to be analysed where each observation is a network. In ecology, this arises when recording species' interactions across multiple different areas or sites. In this setting, the goal is to model the underlying mechanism that generates the multiple network data and to be able to directly compare networks across different spatial or temporal scales. Recent studies have focused on the problem of modelling multiple network data utilising (a) a latent space framework (Gollini and Murphy 2016; Durante et al. 2017; Wang et al. 2019; Nielsen and Witten 2018; Arroyo et al. 2021), (b) a measurement error process (Le et al. 2018; Newman 2018; Peixoto 2018; Mantziou et al. 2024; Young et al. 2022), (c) distance functions (Lunagómez et al. 2020; Kolaczyk et al. 2020; Ginestet et al. 2017; Josephs et al. 2023) and (d) a Stochastic Block Model (SBM) structure (Josephs et al. 2023). Another study utilising an SBM structure for modelling multiplex networks is the study of Amini et al. (2024). Multiplex networks are different to multiple networks as the former refers to networks observed at different layers with edges at each layer having a different interpretation of relation, while in multiple network data, the edges express the same type of relation across networks. In Amini et al. (2024), the authors propose a hierarchical Stochastic Block Model (SBM) to recover communities of nodes at different network layers. However none of these models explicitly consider networks' cyclical properties in their formation, making it difficult to determine the underlying processes that sustain species interactions in the wild, or to measure how they might differ when drivers change. Recent work utilising subgraph counts to test whether networks arise from a given distribution (Maugis et al. 2020) highlights how network properties, such as cycles, can be valuable in the analysis of network populations.

Distance-based models offer a way to encode networks' cycle information by incorporating this information in a metric measuring similarity between networks. There are a multitude of ways to define a similarity measure between different networks and for a review of these see Donnat and

Holmes (2018). However, none of these metrics explicitly consider cycles when measuring network dissimilarity.

In this article we propose a distance-based model for multiple network data that explicitly utilises the cycle information in the distance metric. Specifically, the metric we propose involves counting the number of uncommon cycles between the two networks, denoted as the symmetric difference, and combining this with the Hamming distance (both defined in the next section). The resulting metric is denoted as the Hamming-Symmetric difference (HS) distance metric. Enumerating cycles within a network is a computationally intensive task. To deal with computational challenges in detecting large cycles, in this study we consider only directed cycles formed by three nodes, i.e. directed triangles. From an ecological perspective, three-node motifs are of interest as they indicate where on the transitive to intransitive continuum a species triad falls, with a directed triangle representing intransitive competition. There is thus an ecological, as well as mathematical, interest in studying directed triangles in network data.

We adopt a Bayesian approach and utilise the Spherical Network Family (SNF) of models (Lunagómez et al. 2020) to make posterior inferences as this gives us the flexibility to specify the distance metric of our choice, which in our setting is the HS distance. However, the computational challenges associated with fitting the SNF model are significant, with the model having an intractable normalising constant, which is a sum over the space of graphs. For a detailed review on methods for intractable distributions see Park and Haran (2018). Notably, for an directed network with  $n$ -nodes there are  $2^{n(n-1)}$  possible networks, which means that even for a moderate-sized network with  $n = 20$  nodes there will be more than  $2.46 \times 10^{114}$  network configurations which are not practically possible to enumerate.

The methodology proposed in Lunagómez et al. (2020) utilises a diffusion distance metric and the auxiliary variable implementation based on Møller et al. (2006) to deal with the double intractability problem. The choice of the diffusion distance metric in Lunagómez et al. (2020) is motivated by a neuroscience application, as diffusion distance can capture differences between networks with respect to how messages propagate through brain regions. However, for our ecological application, the specification of the diffusion distance metric would hinder ecological interpretation as it would not be clear how to translate message propagation in this setting. In our framework, we are interested in local changes in the structure of the networks with respect to edge flips and cycles differences rather than the global changes captured by the diffusion distance. The specification of our proposed HS distance metric allows for capturing such properties, however, the auxiliary variable technique formulated in Lunagómez et al. (2020) does not result in satisfactory performance for making posterior inferences in our setting.

We thus develop an alternative computational framework to make posterior inferences through approximating the normalising constant using an Importance Sampler. This was inspired by an approach taken in Vitelli et al. (2017), albeit in a different setting with less computational challenges. The resulting modelling framework performs significantly better in making posterior inferences. More details are given in Section 5.3. We further evaluate our approach on field data of competitive interactions between fish species at various reefs in the Indo-Pacific Ocean.

While the seminal approach taken in Lunagómez et al. (2020) has opened up exciting possibilities for developing interpretable Bayesian models for multiple network data, and which our modelling framework is based on, we note a related earlier study Banks and Carley (1994), which provides one of the first approaches in the literature for modelling multiple network data. Banks and Carley (1994) propose a model with the same functional form to the SNF model presented in Lunagómez et al. (2020), and consider the Hamming distance metric to make inferences for multiple network data. The key differences with Lunagómez et al. (2020) are (i) the inferential framework in Banks and Carley (1994) is formulated for the Hamming distance and variants of it while (Lunagómez et al. 2020) offer an inferential framework that allows the practitioner to specify a distance metric of their choice and (ii) Banks and Carley (1994) do not account for uncertainty quantification as in Lunagómez et al. (2020) who address this with a fully Bayesian framework.

Our key contributions in this paper are thus three-fold. First, we propose a novel network distance metric, namely the HS distance, that has not been considered in the network literature. Second, we develop and implement a novel Markov Chain Monte Carlo (MCMC) scheme to make posterior inferences from the Spherical Network Family (SNF) model for multiple network data under the proposed HS distance metric. Specifically, we introduce an Importance Sampling (IS) step to approximate the SNF model's intractable normalising constants within a Metropolis-Hastings (MH) algorithm. Third we utilise the modelling framework to infer cycle properties from a group of ecological networks studying aggressive interactions between species of fish.

The remainder of this article is organised as follows. Section 2 briefly reviews relevant fundamental network concepts, Section 3 describes our proposed metric as well as the ecological application that motivated its derivation. In Section 4 we give an overview of the SNF model and how Lunagómez et al. (2020) address the problem of the intractable normalising constant. In Section 5 we present how we modify the computations to make posterior inferences for the SNF model and deal with the MCMC mixing issue, along with simulation experiments for evaluating the performance of our method. Section 6 applies our modelling framework to ecological data, specifically to quantify aggressive interactions

between coral-eating reef species of fish. Finally Section 7 ends with some concluding remarks.

## 2 Relevant network properties and preliminaries

We represent a directed graph by  $\mathcal{G} = (V, E)$ , with  $V = \{1, \dots, n\}$  denoting the set of  $n$  nodes and  $E \subseteq \mathcal{E}_n$  denoting the set of edges in  $\mathcal{G}$ , with  $\mathcal{E}_n = \{(i, j) \mid i, j \in V\}$ . Directed networks have ordered edges, such that  $(i, j)$  is distinct to  $(j, i)$ . We note here that in this paper, we focus only on directed networks, however, our framework is easily adaptable to the simpler setting of undirected networks. We use an  $n \times n$  matrix, namely the adjacency matrix, to represent the presence and absence of edges in graph  $\mathcal{G}$ . Thus, the  $(i, j)^{th}$  element of the adjacency matrix for a graph with binary edges is,

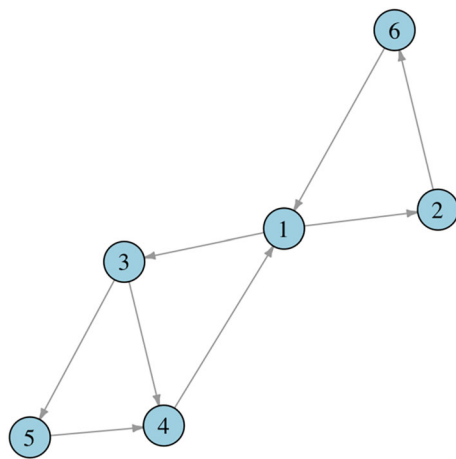
$$A_{\mathcal{G}}(i, j) = \begin{cases} 1, & \text{if an edge occurs from node } i \text{ to node } j, \\ 0, & \text{otherwise.} \end{cases} \quad (1)$$

By  $\mathcal{G}_1, \dots, \mathcal{G}_N$  we represent a population of  $N$  directed graphs, with corresponding adjacency matrices  $A_{\mathcal{G}_1}, \dots, A_{\mathcal{G}_N}$ . We further assume that the networks in the population have no self-loops, and share the same set of  $n$  labelled nodes suggesting that the rows/columns of the adjacency matrices adhere to the same order. We represent the space of graphs with  $n$  nodes by  $\{\mathcal{G}_{|n|}\}$ , such that  $\{\mathcal{G}_{|n|}\} = \{\mathcal{G} = (V, E) : |V| = n\}$ . Thus, the size of the space of directed, with no self-loops graphs is  $|\{\mathcal{G}_{|n|}\}| = 2^{n(n-1)}$ .

A way to quantify similarities among networks is through the use of distance metrics which we denote by  $d_{\mathcal{G}}(\cdot, \cdot)$ . Two main types are: (a) structural distances that aim to capture similarities on edge-specific local properties of the graphs, and (b) to spectral distances that aim to capture similarities with respect to global properties of the graphs using a spectral representation (Donnat and Holmes 2018). A well-known structural distance metric is the *Hamming distance*, that counts the not in common edges and non-edges between two graphs  $\mathcal{G}_k$  and  $\mathcal{G}_l$  for  $k, l \in \{1, \dots, N\}$ . The unnormalised Hamming distance between  $\mathcal{G}_k$  and  $\mathcal{G}_l$  is defined as:

$$d_H(A_{\mathcal{G}_k}, A_{\mathcal{G}_l}) = \sum_{i,j} |A_{\mathcal{G}_k}(i, j) - A_{\mathcal{G}_l}(i, j)|. \quad (2)$$

Networks are objects that can exhibit complex structures, thus the derivation of network properties such as the degree distribution is important for evaluating their characteristics. Network cycles are known to be crucial in revealing information about their topology (Maugis et al. 2017). We



**Fig. 1** Example of graph with directed cycles  $\{1 - 2 - 6 - 1\}$ ,  $\{1 - 3 - 5 - 4 - 1\}$  and  $\{1 - 3 - 4 - 1\}$

acknowledge that terminology about motifs is not universal, and thus we use the following toy network to clarify what we mean by directed triangles which is the focus of our study. Firstly, a *directed cycle* in a directed network is a sequence of connected nodes in which the only repeated nodes are the first and the last node in the sequence. An illustrative example of an directed graph with three directed cycles is presented in Figure 1. We note here that  $\{3, 5, 4\}$  does not form a directed cycle since there is an edge  $(3, 4)$  rather than  $(4, 3)$ . The sequence of nodes  $\{3, 5, 4\}$  and  $\{1, 3, 4\}$  form triangles. In this study, we consider only directed cycles that form triangles (e.g.  $\{1, 3, 4, 1\}$ ), namely *directed triangles*.

### 3 Ecological Application and proposed metric

Data have been collected on aggressive interactions between butterflyfish (genus *Chaetodon*) on different coral reefs, across the Indo-Pacific region (Keith et al. 2018). We use a network representation, where nodes represent fish species and edges represent aggressive encounters.

In Ecology, it is often of interest to identify competition structures among species that share resources, which enable particular ecological dynamics to be inferred e.g., hierarchical versus intransitive competition. Changes in competitive interactions can have broader effects, altering population dynamics, community structure, ecosystem function (Ecological Competition 2022; Mohd 2019; Grether et al. 2017; Kinlock 2021). In particular, the formation of directed cycles in graphs representing aggressive interactions among species point to intransitive competition patterns which are of particular interest ecologically (Sokhn et al. 2012; Koutrouli et al. 2020). Hence, a research question arising here is the following: How can we best use cyclical properties of networks,

together with other network properties, to jointly analyse multiple network data, when the applied research questions pertain to directed cycles? We develop a Bayesian modelling and inferential framework to address this.

An appealing way to answer this research question is with the SNF model (Lunagómez et al. 2020) that infers a representative network in the population, determined through a user-specified distance metric. In addition, the SNF model involves a dispersion parameter that quantifies the level of dissimilarity between the network data and the network representative, with respect to the specified metric. The flexibility in the choice of the distance metric is a key motivating factor for developing a SNF model to analyse the data.

As our interest lies in cycles formed in the network data, we propose a measure that captures information about dissimilarities between the cycles of two networks. Specifically, we propose a Hamming-Symmetric difference (HS) distance metric consisting of two parts:

1. The Hamming distance counting not in common edges/non-edges between two graphs.
2. The symmetric difference between the cycles formed in two graphs, i.e. counting the number of not in common cycles in two graphs.

Hence, a mathematical representation of the constructed distance metric for two graphs  $\mathcal{G}_k, \mathcal{G}_l$  for  $k, l \in \{1, \dots, N\}$  is,

$$d_{\text{HS}}(\mathcal{G}_k, \mathcal{G}_l) = d_{\text{H}}(A_{\mathcal{G}_k}, A_{\mathcal{G}_l}) + \lambda \cdot |C_{\mathcal{G}_k} \Delta C_{\mathcal{G}_l}|, \quad (3)$$

where  $d_{\text{H}}(\cdot, \cdot)$  denotes the Hamming distance,  $C_{\mathcal{G}_i}$  denotes the directed cycles in graph  $i$ ,  $\Delta$  indicates the symmetric difference and  $\lambda \in \mathbb{R}$  is a weighting factor. In Supplementary material Section 1, we show that HS is a distance metric. Under this construction, we encode information about dissimilarities in the structure of the networks, with respect to both their edges and cycles. The tuning of the  $\lambda$  parameter corresponds to how much influence we allow the symmetric difference to have on the total distance. In the rest of this article, we assume  $\lambda$  to be equal to 1, suggesting equal importance between the Hamming and the Symmetric difference distance.

The specification of the HS distance metric for the SNF model induces significant challenges when adopting the MCMC framework proposed by Lunagómez et al. (2020) to make posterior inferences with the SNF model. Notably the mixing of chains is very poor with acceptance rates close to zero, as illustrated in Figure 3 in Section 5.3. This motivated us to develop an alternative computational framework to make posterior inferences with the SNF model and details are given in Section 5.



## 4 Overview of the SNF model

In this section we provide a brief overview of the SNF model proposed in Lunagómez et al. (2020), and why it is a compelling model to consider for this setting. We also highlight some shortcomings with the current implementation which limits its usefulness in our setting.

### 4.1 Motivation and Model formulation

Lunagómez et al. (2020) develop a model for network data inspired by the form of a Normal distribution. Specifically, they assume an underlying mean network representing the network population and a dispersion parameter denoting the variation of the networks about this mean. They express the mean network in terms of a Fréchet mean, as seen in the studies of Ginestet et al. (2017) and Kolaczyk et al. (2020), and the dispersion parameter in terms of an entropy. Under this construction, they obtain the probabilistic mechanisms that generate data sets of multiple network data which they denote the SNF model.

Specifically, if we assume we have a population of directed and unweighted graphs  $\mathcal{G}_1, \dots, \mathcal{G}_N$  then the joint distribution characterised by the SNF model is given by,

$$P(A_{\mathcal{G}_1}, \dots, A_{\mathcal{G}_N} | A_{\mathcal{G}^m}, \gamma) = \frac{1}{Z(A_{\mathcal{G}^m}, \gamma)^N} \exp \left\{ -\gamma \cdot \sum_{i=1}^N \phi(d_{\mathcal{G}}(A_{\mathcal{G}_i}, A_{\mathcal{G}^m})) \right\}, \quad (4)$$

where  $\mathcal{G}^m$  is the Frechét mean,  $d_{\mathcal{G}}(\cdot, \cdot)$  is a distance metric,  $\gamma > 0$  is the dispersion,  $\phi(\cdot) > 0$  is a monotone increasing function and the model partition function is

$$Z(A_{\mathcal{G}^m}, \gamma) = \sum_{A_{\mathcal{G}} \in \{\mathcal{G}_{|n|}\}} \exp\{-\gamma \cdot \phi(d_{\mathcal{G}}(A_{\mathcal{G}}, A_{\mathcal{G}^m}))\}, \quad (5)$$

where  $\{\mathcal{G}_{|n|}\}$  is the space of  $n$ -node networks. The parameters  $\mathcal{G}^m$ , and  $\gamma$  can thus be seen to relate to the mean and precision parameters in a Normal distribution.

Lunagómez et al. (2020) show the Centered Erdős-Rényi (CER) model is a special case of a SNF model when the Hamming distance metric is used. Under the CER model a population of networks is generated by perturbing the edges of a centroid network  $\mathcal{G}^m$  using a Bernoulli distribution with probability  $\alpha$ , as follows:

$$A_{\mathcal{G}}(i, j) | (A_{\mathcal{G}^m}(i, j), \alpha) = | A_{\mathcal{G}^m}(i, j) - Z(i, j) |, \quad (6)$$

where  $\mathcal{G}^m$  is the Frechét mean and  $Z(i, j)$ 's are *iid*  $\text{Ber}(\alpha)$ , with  $0 < \alpha < 0.5$ . The joint distribution of a population of

directed and unweighted  $\mathcal{G}_1, \dots, \mathcal{G}_N$  graphs is then

$$P(A_{\mathcal{G}_1}, \dots, A_{\mathcal{G}_N} | A_{\mathcal{G}^m}, \alpha) = \prod_{i=1}^N \alpha^{d_H(A_{\mathcal{G}_i}, A_{\mathcal{G}^m})} \cdot (1 - \alpha)^{n(n-1) - d_H(A_{\mathcal{G}_i}, A_{\mathcal{G}^m})} \quad (7)$$

where  $d_H(\cdot, \cdot)$  denotes the Hamming distance metric and  $n$  is the number of nodes.

To make inferences, the authors adopt a Bayesian approach. A prior distribution for  $\gamma$  is specified with support on  $\mathbb{R}^+$ . The prior choice and support is strongly related to the specified distance metric. A prior distribution for the network representative  $A_{\mathcal{G}^m}$  is specified with the same functional form as that of the SNF model. The priors for the parameters of the CER model are specified in a similar manner, with the prior for the representative having the functional form of the CER model. The prior for  $\alpha$  in the CER model requires support on  $(0, 0.5)$  with a scaled Beta distribution on  $(0, 0.5)$  proposed.

### 4.2 Addressing the intractable normalising constant

Lunagómez et al. (2020) make posterior inferences using a MCMC scheme to draw samples from the posterior distribution based on a Metropolis-Hastings (MH) algorithm. However, the normalising constant of the SNF model,  $Z(A_{\mathcal{G}^m}, \gamma)$ , depends on the parameters of the model. Thus, the normalising constants do not cancel in the Metropolis-Hastings ratio.

To tackle the intractable normalising constants, Lunagómez et al. (2020) apply the Auxiliary Variable technique presented in Møller et al. (2006). Notably, Møller et al. (2006) consider a likelihood of the form,

$$P(y | \theta) = \frac{q_{\theta}(y)}{Z(\theta)}, \quad (8)$$

where  $\theta$  denotes the model parameter,  $y$  represents the data,  $q_{\theta}(y)$  the unnormalised density, and  $Z(\theta)$  is an intractable normalising constant that depends on  $\theta$ .

They propose the use of an auxiliary variable  $x$  that has the same support as that of  $y$ , with density  $f(x | \theta, y)$  to obtain an unbiased estimator of  $Z(\theta)$ . In light of Importance Sampling, under Møller et al. (2006)  $Z(\theta)$  can be written as,

$$Z(\theta) = \mathbb{E} \left[ \frac{q(x | \theta)}{f(x | \theta, y)} \right], \quad (9)$$

where the expectation is taken with respect to the density of the auxiliary variable  $x$ ,  $f(x | \theta, y)$ . In this regard, they propose sampling  $x$  from  $P(\cdot | \theta)$  as seen in equation (8) and use the approximation,

$$Z(\theta) \approx \frac{q(x | \theta)}{f(x | \theta, y)}. \quad (10)$$

Thus, they can substitute the normalising constant  $Z(\cdot)$  by its unbiased estimator  $q(x | \cdot)/f(x | \cdot, y)$  in the MH acceptance ratio. In this respect,  $q(x | \cdot)$  and  $f(x | \cdot, y)$  are evaluated in each MCMC iteration at the auxiliary variable  $x$  drawn from the proposal  $P(\cdot | \theta)$ .

The formulation of the Auxiliary Variable Method in the case of the SNF model involves the simulation of a set of auxiliary variables  $\mathcal{G}^*$ , defined on the same state space as the network data  $\{\mathcal{G}_i\}_{i=1}^N$ . Lunagómez et al. (2020) exploit the probabilistic mechanism of the CER model to specify the conditional density  $f(A\mathcal{G}_1^*, \dots, A\mathcal{G}_N^* | \{\mathcal{G}_i\}_{i=1}^N, A\mathcal{G}^m, \tilde{\alpha})$  of the auxiliary network variables  $\mathcal{G}_1^*, \dots, \mathcal{G}_N^*$ . Thence, in each iteration of the MH algorithm a new state of both the model parameters and the auxiliary network variables will be proposed, with the latter sampled from a proposal distribution that has the same functional form as the likelihood. Under this formulation, the normalising constants cancel in the MH ratio. For a more detailed description of this MH algorithm see Lunagómez et al. (2020).

A main challenge in implementing the Auxiliary Variable Method for the SNF model is the slow mixing of the chain for  $\gamma$ , as seen in Figure 3 in Section 5.3. Notably, we see a very low acceptance rate and thus poor mixing. Depending on the distance metric choice, this issue is apparent even for small network sizes.

The occurrence of this phenomenon can be attributed to the discrepancy between the likelihood, the SNF model, and the choice of auxiliary density, the CER model. Depending on the choice of the distance metric for the SNF model, this discrepancy can increase leading to a bad mixing or, in some cases, the chain not exploring the state space at all.

The poor mixing makes this impractical to consider for our setting and motivates our development of an alternative strategy to approximate the normalising constant. The proposed approach greatly improves performance of the MCMC, allowing it to be applied to similar sizes of networks present in the ecological data set.

## 5 Proposed Bayesian inference framework for the SNF model using Importance Sampling

To overcome shortcomings of the Auxiliary Variable approach we develop an alternative method to approximate the intractable normalising constant. Specifically, we formulate an Importance Sampling step within our MCMC equivalent to Ratio Importance Sampling (Chen and Shao 1997). We were motivated by Vitelli et al. (2017) who also use Importance Sampling to make Bayesian inference from the

Mallow's model (Mallows 1957), a common model for analysing rank data with the same functional form as the SNF model. Frequentist inference may also be possible, with Mardia and Dryden (1999) developing this for the Watson model that also has the same functional form as the SNF model.

A key difference between the Mallow's model for rank data and the SNF model is that the normalising constant in the latter involves both the representative network and the dispersion parameter, while for the Mallow's model the normalising constant depends only on the dispersion parameter, for right-invariant distance metrics considered in Vitelli et al. (2017). This allows an off-line approximation of the normalising constant through IS, using a pseudo-likelihood approximation of the target distribution.

Graphs are more complex objects than rank data, due to diverse structures they exhibit such as the formation of communities and motifs, as well as other topological structures revealed by their spectral decomposition. For networks, the right-invariance property does not hold for the majority of distance functions, with different properties governing rank and network data. Thus, approximating the normalising constant of the SNF model is a more challenging scenario. Unlike Vitelli et al. (2017), we formulate an Importance Sampler within our MCMC to give a good approximation to the normalising constant.

### 5.1 Formulation of IS step for the SNF model

The normalising constant of the SNF model has the following form,

$$Z(A\mathcal{G}^m, \gamma) = \sum_{A\mathcal{G} \in \{\mathcal{G}_{|n|}\}} \exp\{-\gamma \cdot \phi(d_{\mathcal{G}}(A\mathcal{G}, A\mathcal{G}^m))\}, \quad (11)$$

this involves computing a sum over the space of  $n$ -node graphs,  $\{\mathcal{G}_{|n|}\}$ . Even for modest  $n$ , this sum is impractical to compute. Instead, using ideas from Importance Sampling (Robert and Casella 2013) we can rewrite the sum as

$$\begin{aligned} & \sum_{A\mathcal{G} \in \{\mathcal{G}_{|n|}\}} \exp\{-\gamma \cdot \phi(d_{\mathcal{G}}(A\mathcal{G}, A\mathcal{G}^m))\} \\ &= \sum_{A\mathcal{G} \in \{\mathcal{G}_{|n|}\}} \frac{\exp\{-\gamma \cdot \phi(d_{\mathcal{G}}(A\mathcal{G}, A\mathcal{G}^m))\}}{g(A\mathcal{G})} g(A\mathcal{G}) \\ &= \mathbb{E}_g \left[ \frac{\exp\{-\gamma \cdot \phi(d_{\mathcal{G}}(A\mathcal{G}, A\mathcal{G}^m))\}}{g(A\mathcal{G})} \right], \end{aligned} \quad (12)$$

which can then be approximated by drawing a sample of networks  $\mathcal{G}_1, \dots, \mathcal{G}_K$  from an Importance Sampling (IS) proposal density  $g$  and calculating,

$$\hat{Z}(A\mathcal{G}^m, \gamma) \approx \frac{1}{K} \sum_{k=1}^K \frac{\exp\{-\gamma \cdot \phi(d_{\mathcal{G}}(A\mathcal{G}_k, A\mathcal{G}^m))\}}{g(A\mathcal{G}_k)}. \quad (13)$$

One advantage of the IS method is the flexibility with specifying the IS density. In this regard, choices of distributions that are easy to sample from are preferred (Robert and Casella 2013). In our problem, a natural choice of the IS density is the distance-based CER model for two main reasons, (i) the CER model is a member of the Spherical Network Family (SNF) of models (Lunagómez et al. 2020), and (ii) sampling network data from the CER model is quick, thus will result in a less computationally intensive MCMC algorithm. To sample networks from the CER model, we perturb the edges of the centroid  $\tilde{A}_{\mathcal{G}^m}$  using Bernoulli noise with probability  $\tilde{\alpha}$ , as per equation (6).

Thus, the estimator in (13) takes the following form under the CER IS density:

$$\hat{Z}(A_{\mathcal{G}^m}, \gamma) \approx \frac{1}{K} \sum_{k=1}^K \frac{\exp\{-\gamma \cdot \phi(d_{\mathcal{G}}(A_{\mathcal{G}_k}, A_{\mathcal{G}^m}))\}}{\tilde{\alpha}^{d_{\mathcal{H}}(A_{\mathcal{G}_k}, \tilde{A}_{\mathcal{G}^m})} (1 - \tilde{\alpha})^{n(n-1) - d_{\mathcal{H}}(A_{\mathcal{G}_k}, \tilde{A}_{\mathcal{G}^m})}}, \quad (14)$$

where  $\{A_{\mathcal{G}_k}\}_{k=1}^K$  are networks sampled from the CER model with parameters  $\tilde{\alpha}$  and  $\tilde{A}_{\mathcal{G}^m}$ .

We determine  $\tilde{\alpha}$  and  $\tilde{A}_{\mathcal{G}^m}$  by fitting the data to the CER model to obtain the posterior mean of  $\tilde{\alpha}$  and posterior mode of  $\tilde{A}_{\mathcal{G}^m}$ . In this way, we encode information about the data that may allow a better approximation of the normalising constant.

## 5.2 MCMC scheme with IS step

We now describe our computational framework to obtain posterior draws for the SNF model parameters. As seen in Lunagómez et al. (2020), the joint posterior distribution of the centroid  $A_{\mathcal{G}^m}$  and the dispersion parameter  $\gamma$  can be expressed as

$$P(A_{\mathcal{G}^m}, \gamma | A_{\mathcal{G}_1}, \dots, A_{\mathcal{G}_N}) \propto \frac{1}{Z(A_{\mathcal{G}_0}, \gamma_0)} \exp\{-\gamma_0 \phi(d_{\mathcal{G}}(A_{\mathcal{G}^m}, A_{\mathcal{G}_0}))\} P(\gamma | \alpha_0) \cdot \frac{1}{Z(A_{\mathcal{G}^m}, \gamma)^N} \exp\{-\gamma \sum_{i=1}^N \phi(d_{\mathcal{G}}(A_{\mathcal{G}_i}, A_{\mathcal{G}^m}))\}. \quad (15)$$

We follow a largely similar scheme to Lunagómez et al. (2020) to make inferences, using Metropolis-Hastings to sample from the joint posterior of the parameters. However, to overcome the double-intractability problem, we approximate the normalising constant within each iteration of the MCMC using the estimator obtained through Importance Sampling, different to the Auxiliary Variable Method adopted by Lunagómez et al. (2020). Notably, we obtain posterior draws from the target distribution in Equation (15),

after substituting the normalising constant in the likelihood with its estimate in equation (14).

To obtain posterior draws for the parameters  $A_{\mathcal{G}^m}$  and  $\gamma$ , we follow a similar scheme to Lunagómez et al. (2020). Details are given in Supplementary material Section 2.

Algorithm 1 sketched below illustrates the MH algorithm with IS step.

### Algorithm 1: Metropolis-Hastings Algorithm with IS step

**Data:**  $A_{\mathcal{G}_1}, \dots, A_{\mathcal{G}_N}$  **Hyperparameters:**  
 $A_{\mathcal{G}_0}, \gamma_0, \alpha_0, \tilde{\alpha}, \tilde{A}_{\mathcal{G}^m}$

**Initialisation:** Randomly generate  $\gamma^{(0)}$  and  
 $A_{\mathcal{G}^m}^{(0)} \sim \text{Bernoulli}(\sum_{i=1}^N A_{\mathcal{G}_i} / N)$

**for**  $i \leftarrow 1$  **to**  $M$  **do**

**MH step with a mixture of kernels:** Update  $A_{\mathcal{G}^m}$  or  $\gamma$

Sample  $v \sim \text{Multinomial}(\xi_1, \dots, \xi_L)$

Depending on the value of  $v$  propose

$A_{\mathcal{G}^m}^{(i)} \sim q(A_{\mathcal{G}^m}^{(i)} | A_{\mathcal{G}^m}^{(i-1)})$

or  $\gamma^{(i)} \sim q(\gamma^{(i)} | \gamma^{(i-1)})$

**Draw new IS sample of networks:**

$A_{\mathcal{G}_1}^{IS(i)}, \dots, A_{\mathcal{G}_K}^{IS(i)} \sim \text{CER}(\tilde{\alpha}, \tilde{A}_{\mathcal{G}^m})$ .

**Estimate Z:** Use equation (14) to estimate normalising constant in posterior  $P(A_{\mathcal{G}^m}^{(i)}, \gamma^{(i)} | A_{\mathcal{G}_1}, \dots, A_{\mathcal{G}_N})$

**Calculate MH ratio:**

$$r = \min\left(1, \frac{P(A_{\mathcal{G}^m}^{(i)}, \gamma^{(i)} | A_{\mathcal{G}_1}, \dots, A_{\mathcal{G}_N}) \cdot q(A_{\mathcal{G}^m}^{(i-1)}, \gamma^{(i-1)} | A_{\mathcal{G}^m}^{(i)}, \gamma^{(i)})}{P(A_{\mathcal{G}^m}^{(i-1)}, \gamma^{(i-1)} | A_{\mathcal{G}_1}, \dots, A_{\mathcal{G}_N}) \cdot q(A_{\mathcal{G}^m}^{(i)}, \gamma^{(i)} | A_{\mathcal{G}^m}^{(i-1)}, \gamma^{(i-1)})}\right)$$

$u \sim \text{Bernoulli}(r)$

**if**  $u=1$  **then**

Accept proposals  $A_{\mathcal{G}^m}^{(i)}, \gamma^{(i)}$

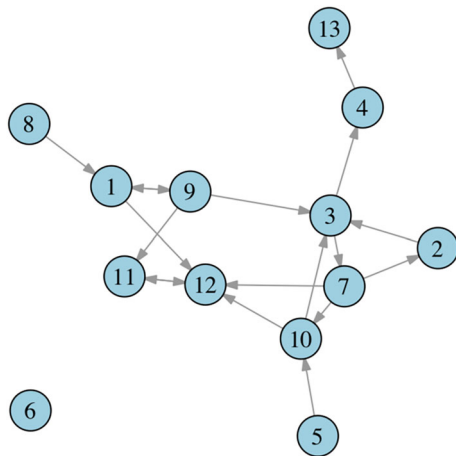
**else**

Reject proposals  $A_{\mathcal{G}^m}^{(i)}, \gamma^{(i)}$

**end**

**end**

We randomly generate an initial centroid network  $\mathcal{G}^{m(0)}$  by sampling edges independently from a Bernoulli distribution with probabilities equal to the average adjacency matrix obtained from the observed network population  $\sum_{i=1}^N A_{\mathcal{G}_i} / N$ . Thus, we assist our MCMC with a meaningful network initialisation using information from the observed network population. In our implementation, we initialise  $\gamma^{(0)}$  at 0.1. In practice, any real positive can be specified, ideally upper bounded by the value of  $\gamma$  for which the average distance of the networks from the centroid is close to 0 (see Figure 4). For the mixture of kernels we consider proposals imposing both moderate and more drastic changes in the current values of the parameters. Our investigation suggests that inferences are not sensitive to moderate changes in the probabilities. The prior specification for the centroid and the



**Fig. 2** Simulated centroid with 13 nodes

dispersion is performed as suggested in Lunagómez et al. (2020).

### 5.3 Addressing the MCMC chain mixing issue

In this section, we illustrate the improvement in the MCMC chain mixing using the IS step compared to the auxiliary variable technique used in Lunagómez et al. (2020), for the HS distance metric.

In this simulation experiment, we consider network size, population size and parameter values similar to that of the ecological application. In particular, we simulate a population of  $N = 13$  networks with  $n = 13$  nodes, under the scenario of a 13-node centroid with density approximately 0.1 (Figure 2) and  $\gamma = 1.2$ . Similarly to Lunagómez et al. (2020), we simulate from the SNF model with this parameter specification using an MH algorithm with target distribution the density of the SNF model as seen in equation (4). In Figure 3, we show the results after running the MCMC using the auxiliary variable technique for 5,000 iterations (left) and our proposed MCMC with IS step for 50,000 iterations (right), with an IS sample of 3000 networks. We observe a drastic improvement in the mixing of the MCMC chain using our proposed MCMC scheme.

The auxiliary variable technique requires sampling from the SNF model using an MH algorithm in each MCMC iteration, resulting in a computationally intensive algorithm. Indicatively, running the MCMC with the auxiliary variable technique for 5,000 iterations in this simulation experiment requires approximately 20 hours. In contrast, our proposed MCMC scheme with IS step does not require running an MH algorithm in each MCMC iteration since we can sample directly from the CER model using (6). In this simulation experiment, running our proposed MCMC with IS step for 50,000 iterations requires approximately 20 hours. Running each MCMC scheme for 5,000 iterations and 50,000 itera-

tions respectively, results in running the two approaches for a comparable amount of time. Thus, our proposed approach not only improves mixing but also substantially improves computation time.

We observe that the posterior region for  $\gamma$  explored by the MCMC is slightly lower than the true value of  $\gamma = 1.2$  specified to simulate the network population. An explanation is possible by examining the EDA violin plots obtained in Figure 4, showing the distribution of the HS distance between the centroid and simulated networks from the SNF model for different  $\gamma$  values. We see when the true value of  $\gamma$  lies in  $(0.9, 1.21)$  the distribution of the distance between the simulated networks and the centroid is similar. Thus, for this regime, changes in the parameter space result in very small changes in the distribution, making the estimation of  $\gamma$  in this regime a more challenging task.

The scope of this experiment is to illustrate the performance of our method and highlight the improvement of the MCMC chain mixing compared to the auxiliary variable technique, in a similar setting to that of the real data application. In the next section, we consider a regime with higher noise and a centroid with more cycles, to further illustrate the performance of our method in recovering  $\gamma$  and the practicality of our proposed approach in recovering cycles compared to the baseline CER model.

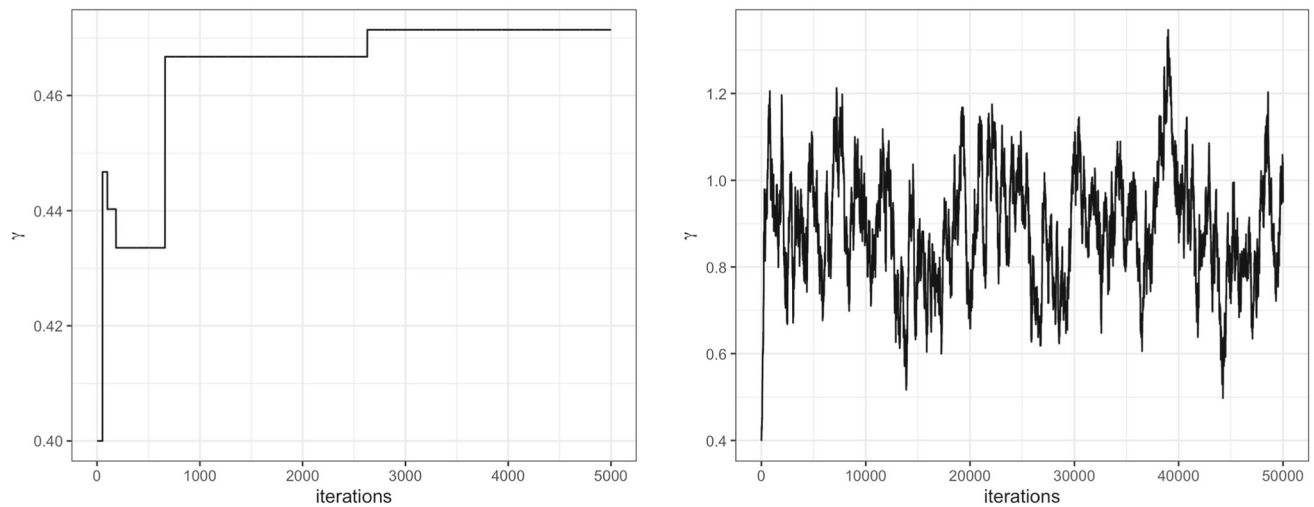
### 5.4 Network population with more cycles and high noise

We consider a simulation regime with a 13-node centroid enclosing 26 directed triangles, and high noise indicated by a smaller size of  $\gamma$ . Notably we specify  $\gamma = 0.2$  for which regime the distribution of distance in its neighbourhood is more distinguishable. We simulate a population of  $N = 100$  networks using the SNF model with HS distance.

Figure 5 shows the traceplot of the posterior draws for  $\gamma$  for 50,000 iterations of the MCMC, and the density of the posterior distribution for  $\gamma$  with the 95% credible interval (blue dashed lines) and the true size of  $\gamma$  (red dashed line). The results suggest accurate recovery of the true  $\gamma$  with posterior mean 0.19 and the true value of  $\gamma$  lying within the 95% credible interval.

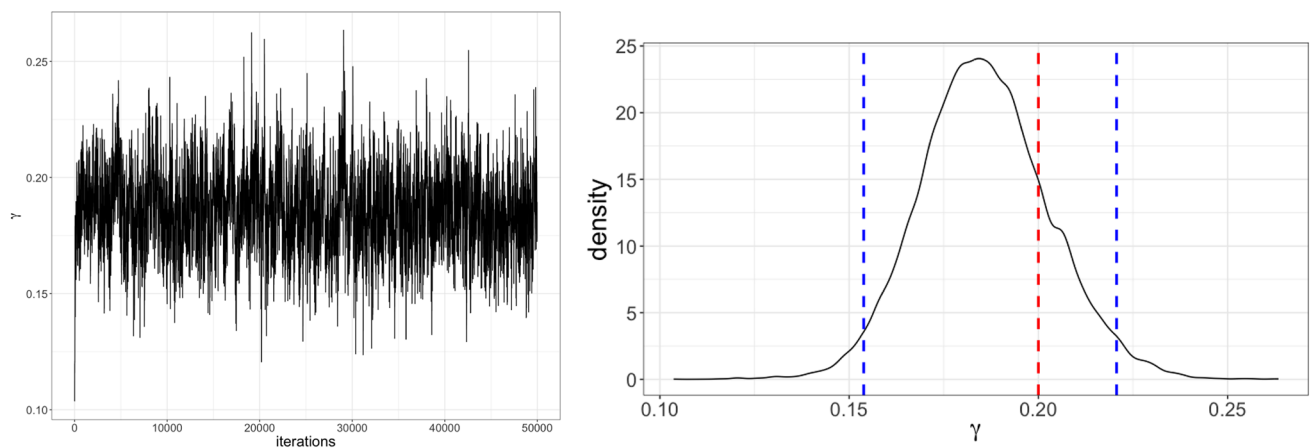
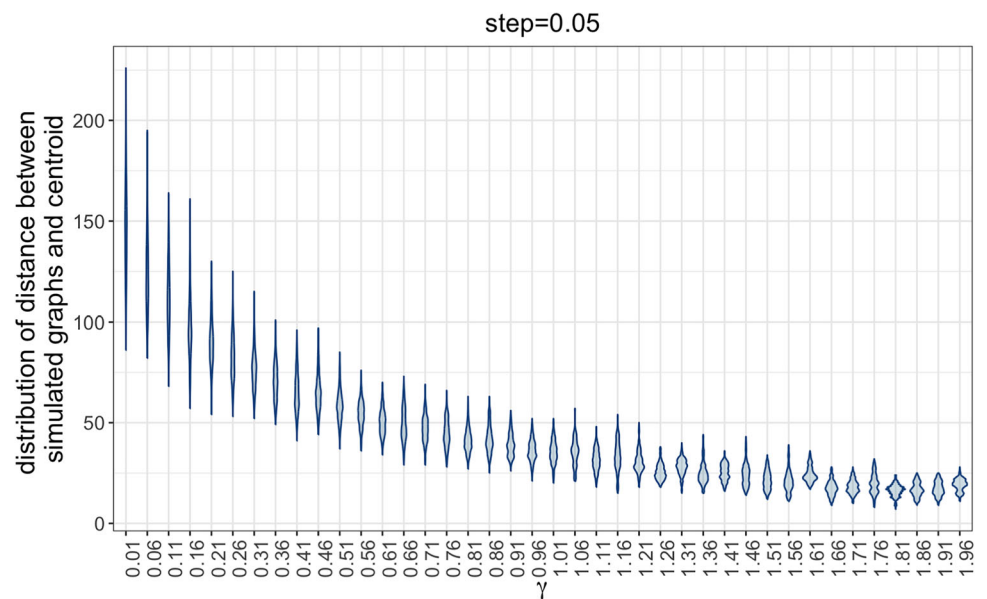
We also examine the ability of our approach in recovering the cycles of the true centroid and compare it to the CER model as a baseline. We summarise the results from the posterior draws for the centroid by calculating the Posterior Inclusion Probability (PIP) of the directed triangles which are present in the true centroid, for the SNF with HS distance and the CER model respectively, as shown in Figure 6. The results shown are for 50,000 MCMC iterations with a burn-in of 1,000 iterations for the SNF model with HS distance, and for 200,000 MCMC iterations with a burn-in of 150,000 iterations for the CER model. We observe that the SNF model



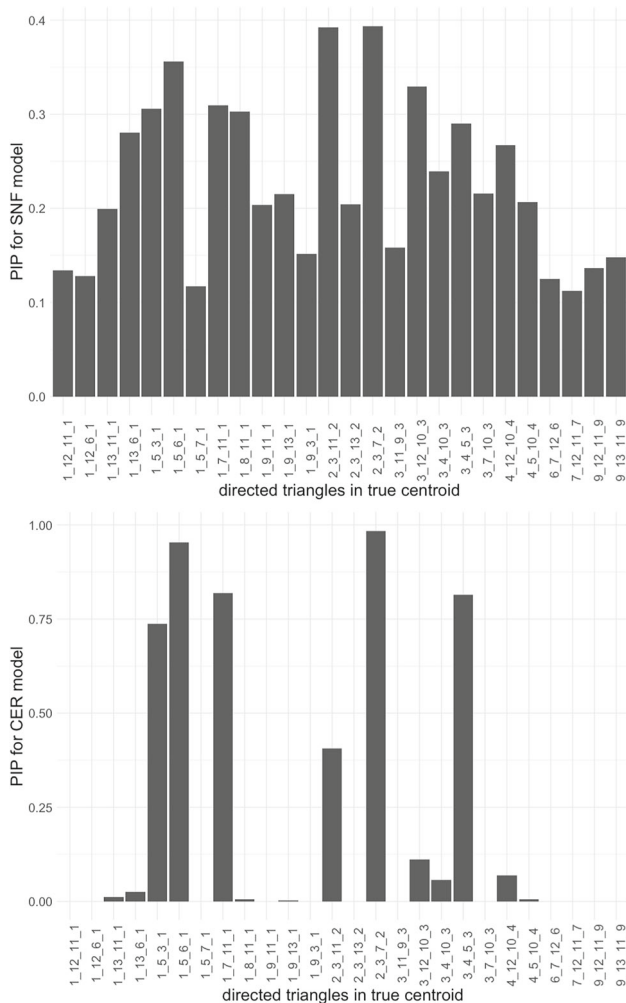


**Fig. 3** MCMC samples for the dispersion parameter  $\gamma$  in the SNF model with HS distance metric using the auxiliary variable method in Lunagómez et al. (2020) (left), versus using our framework utilising Importance Sampling (right), for the same simulated network population

**Fig. 4** Distribution of HS distance between simulated graphs from SNF model and centroid  $\mathcal{G}^m$ , for varying  $\gamma$  values



**Fig. 5** MCMC samples for the dispersion parameter  $\gamma$  in the SNF model with HS distance metric using IS (left), density plot of MCMC samples with blue dashed lines indicating the 95% credible interval and red dashed line indicating the true size of  $\gamma = 0.2$  (right)

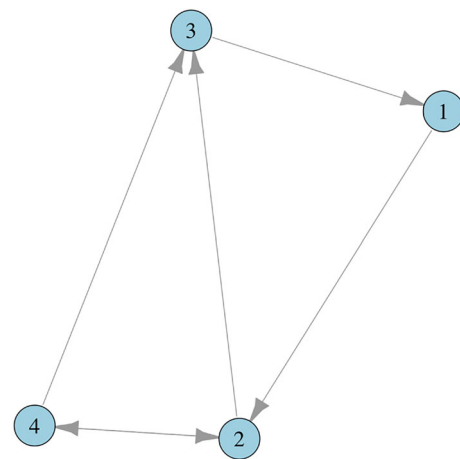


**Fig. 6** Posterior Inclusion Probabilities (PIP) for directed triangles in true centroid for MCMC samples of SNF model with HS distance metric using IS (top), and of CER model (bottom)

with HS distance is able to recover all the directed triangles present in the true centroid with PIP up to 0.4, while the CER model is able to recover only approximately half of the true directed triangles (14 out of 26) with up to 0.98 PIP. This result highlights the key difference between the two models; on the one hand the SNF model with HS distance explicitly accounts for directed cycles but with smaller PIP due to the presence of high noise with respect to directed triangles in the simulated population. On the other hand the directed triangles detected using the CER model are only incidental and an artifact of the frequency of specific edges within the simulated network population.

### 5.5 Empirical evaluation of normalising constant approximation

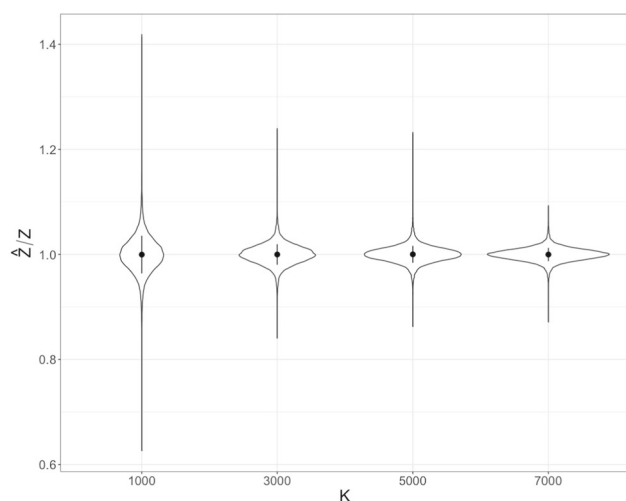
We now evaluate the performance of our IS approximation to the normalising constant in a simulation setting. Specifically,



**Fig. 7** Simulated centroid with 4 nodes

we consider the approximation applied to 4-node directed networks. Here there are  $2^{12} = 4,096$  possible networks, which is a small enough set to calculate the normalising constant exactly under a given model parameterisation in each iteration of the MCMC. This is then used as a benchmark with which to compare the normalising constant approximated through IS.

We simulate a population of  $N = 13$  networks, similarly to the population size in our ecological application, using the SNF model with dispersion  $\gamma = 1$  and 4-node directed network centroid, as shown in Figure 7, enclosing 1 directed triangle. We run our MCMC scheme with IS step for 50,000 iterations, and in each iteration, we also calculate the true normalising constant. To further explore the sensitivity of the approximation to the IS sample size  $K$ , we consider a range of values for  $K = \{1000, 3000, 5000, 7000\}$ . In Figure 8 we present the distribution of the ratio of the estimated normalising constant  $\hat{Z}(A_{G^m}, \gamma)$  and the exact normalising constant  $Z(A_{G^m}, \gamma)$  for varying  $K$  sizes and 50,000 iterations of our MCMC. We observe that the distribution of the ratio has mean (points in Figure 8) and median equal to 1 for all  $K$ , and standard deviation (error bars in Figure 8) of 0.036, 0.019, 0.016 and 0.013 for each  $K$  respectively. The approximation is only marginally sensitive to the size of  $K$  for  $K \geq 3000$ , which justifies the choice of an IS sample  $K = 3000$  networks, to avoid additional computational complexity of cycle detection for large samples of networks in each MCMC iteration. The results indicate that our proposed IS step not only improves the mixing of the MCMC chain (see Section 5.3), but it is also a good approximation with respect to the exact  $Z(A_{G^m}, \gamma)$  even for small  $K = 1000$ .



**Fig. 8** Distribution of ratio  $\hat{Z}(A_G^m, \gamma)/Z(A_G^m, \gamma)$  for 50,000 iterations of our MCMC scheme with IS step for  $K = \{1000, 3000, 5000, 7000\}$ , for 4-node directed networks

## 6 Reef fish aggressive interactions

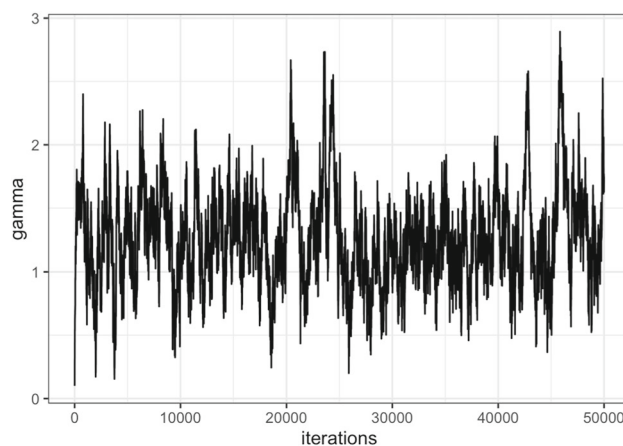
We analyse data collected at various reefs at different regions (Philippines, Bali, Christmas, Iriomote) in the Indo-Pacific ocean (Keith et al. 2018). Each network observation represents the competitive interactions between species of fish. We align the node set across all reefs to comprise the fish species that are in common across all regions, resulting in networks with 13 nodes. The node labels used to represent the species are presented in Table 1. The resulting number of reefs (networks) in our sample is 13, and are visually represented in Supplementary material, Section 3. In this network population, there are three reefs for which directed triangles are observed, as shown in Table 2. We observe that there are no reefs sharing common cycles.

**Table 1** Node labels for species across all regions

species	label
auriga	1
baronessa	2
citrinellus	3
ephippium	4
kleinii	5
lunula	6
lunulatus	7
ornatissimus	8
rafflesii	9
speculum	10
trifascialis	11
unimaculatus	12
vagabundus	13

**Table 2** Triangles in observed networks (reefs)

Reef	Cycle
Jemeluk	2-3-5-2
	2-5-7-2
	2-9-5-2
	3-5-7-3
	3-5-13-3
	5-7-9-5
Lipah	2-11-3-2
Nata	7-12-11-7



**Fig. 9** Traceplot for  $\gamma$  for SNF fitted on fish networks from all regions

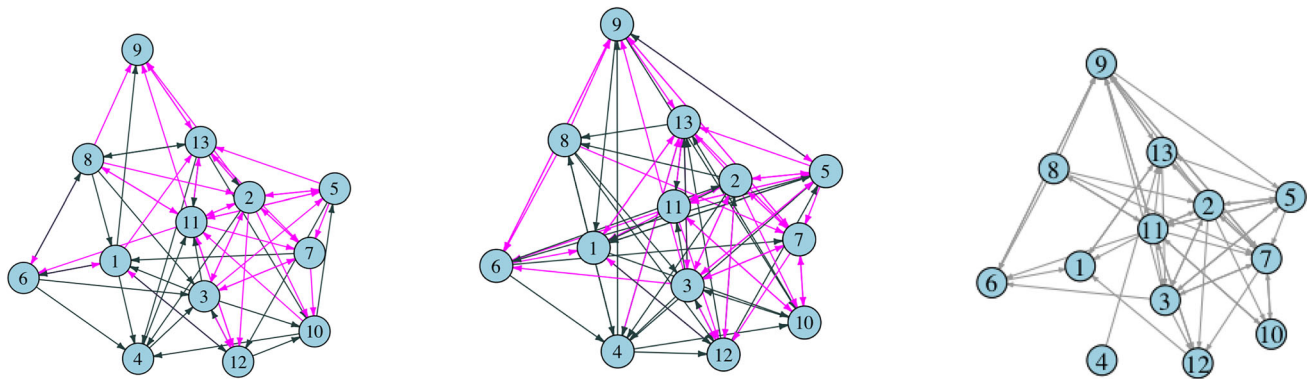
To fit the SNF model with HS distance, we first tune the prior distributions of the parameters and the IS density. We tune the prior for the centroid and dispersion parameter in a similar manner as in Lunagómez et al. (2020). Specifically, we centre the prior for the centroid at the network observation that minimises the distance from the rest of the networks in the sample (Lunagómez et al. 2020). We specify a Gamma prior distribution for the dispersion parameter  $\gamma$  and center it with respect to the average HS distance of the network data from the centroid estimate, where the centroid estimate is obtained by majority vote (connect nodes  $i$  and  $j$  in centroid estimate if the majority of the network population has an edge between  $i$  and  $j$ ). The size of each IS sample is set to  $K = 3000$  networks.

We run our MCMC for 50,000 iterations and obtain summaries of the posterior centroids and associated posterior draws of the dispersion parameter  $\gamma$  of the SNF model. Figure 9 shows a traceplot for the parameter  $\gamma$ .

To investigate our model's efficacy in capturing directed cycle information in the network data, we obtain the 10 most common directed triangles enclosed in the posterior centroids, along with the proportion of the posterior centroid samples containing each cycle, and detect whether these cycles are also observed in the data. In Table 3 we present the 10 most common directed triangles identified in the posterior

**Table 3** Most common directed triangles in posterior draws for centroid  $G^m$ 

SNF model with HS distance directed triangles		
most common directed triangles	proportion of times identified	observed/ inferred
2-3-7-2	0.52	inferred
3-5-7-3	0.49	observed
2-5-11-2	0.49	inferred
2-5-7-2	0.48	observed
2-5-3-2	0.48	inferred
3-9-5-3	0.47	inferred
2-9-7-2	0.46	inferred
3-5-13-3	0.46	observed
2-9-11-2	0.43	inferred
2-3-5-2	0.42	observed

**Fig. 10** Two posterior centroids with highest posterior mass 0.002 (left) and 0.0015 (middle). Pink edges indicate edges also present in the data. Signal to noise ratio (number of edges in common/number of edges inferred) 1.44 and 1.38 respectively. Union of network population (right)

draws for the centroid network after 10,000 iterations burn-in, along with the proportion of times identified and whether the cycle is observed in the network data (observed) or not observed in the network data (inferred).

We observe that 4 out of the 10 directed triangles in the top 10 most common directed triangles are also observed in the real data, with the rest of them enclosing nodes only from the set of nodes observed in the directed triangles present in the network population (Table 2). This indicates that our MCMC algorithm meaningfully accepts posterior centroids with respect to directed triangles observed in the network population. Moreover, there is evidence to suggest that the model is assigning posterior weight to directed triangles based on information in the network data as opposed to simply sampling networks with directed triangles formed by randomly picking 3 unique nodes from amongst those that form the observed directed triangles present in the data (comprising 8 distinct nodes). If it were we would expect each of the  $\binom{8}{3}$  possible directed triangles here to have approximately equal posterior inclusion probabilities of  $1/56$  but the most commonly identified cycles have posterior inclusion probabilities much greater than this.

In Figure 10, we further illustrate the two networks (left and middle graphs) with highest posterior mass. We highlight which edges of these posterior samples are also present in any of the network data in pink. We note that the two posterior centroids, taken from the high posterior mass region, have small posterior mass. This is to be expected when making inferences across a large space of possible graphs coupled with a diverse set of network data. The network at the right in Figure 10 encloses the union of the edges of all observed networks for ease of comparison.

For comparison we fit the CER model to the data, that only considers the Hamming distance. After running the CER model on the network population for 50,000 iterations with a burn-in of 10,000 iterations, we observe that the model is unable to make inferences on directed triangles. Notably, none of the posterior centroids encloses a directed triangle, despite the presence of directed triangles in the network population. This is anticipated for two reasons, (i) as the CER model assumes that the centroid is polluted by Bernoulli noise, increasing the network population results in Bernoulli noise corrupting the cycles in the data rather than preserving them, and (ii) transitivity for Erdős-Rényi models is very low. In contrast, the SNF model explicitly accounts for directed



cycles through the HS distance metric. This finding highlights the importance of our proposed modelling framework when it is of interest to capture the formation of directed cycles in a network population.

## 7 Conclusion

Modelling multiple network data is essential to addressing many applied research questions. In this article we proposed a metric that explicitly incorporates networks' cycles, denoted as the HS distance. We incorporated this metric within the SNF model and developed a computational framework to allow posterior inferences in practical ecological settings, hitherto not possible with the original implementation of the model. We applied our modelling framework to make inferences from ecological data studying aggressive interactions between species of fish and were able to infer cyclical properties that were not possible to detect using a simpler CER model that does not account for cycle information.

While we have shown benefits of our approach in the analysis of ecological studies, cycles are of interest in many other fields such as neuroscience or genetics. The HS distance can thus be an informative measure in many other network applications. A key challenge in these settings may be the larger size of networks (number of nodes). Dealing with cycles is already a substantial computational endeavour, and this issue will be exacerbated in larger networks. Specifically, cycle detection is the main bottleneck for scaling to larger networks, since our approach requires the use of an IS sample of networks and the calculation of cycles within this sample. We addressed this here through restricting the model to only consider directed triangles. This was relevant for the ecological study which focused on intransitive competition, and triadic isomorphs are commonly analysed in ecology due to their ease of interpretation. Considering other strategies to deal with the computational burden of calculating cycles would also be of interest. A possible direction would be to obtain an approximation of the HS distance metric using a machine learning approach.

Another interesting direction for future research would be to consider a model selection framework for the SNF model, which could potentially inform the size of  $\lambda$  in the HS distance metric for a given data example.

The ecological study representing the species interactions also contained edges weights, corresponding to the multiple interactions observed between species of fish. It would be interesting to consider how this feature could be incorporated within our modelling framework. A modification of the Hamming distance in the HS metric would be required to quantify dissimilarities between weighted graphs, with the Frobenius distance an alternative. We would also need to adapt the computational framework to permit sampling of

networks from a weighted space of graphs. This would be necessary for both sampling network representatives in the MCMC as well for approximating the normalising constant through Importance Sampling. There is no straightforward solution to addressing this, but if a model could be developed and implemented practically, it would open up the possibility to model a wide range of weighted networks previously not possible with existing methods.

There have been various techniques developed to approximate intractable normalising constants. We implemented an IS that offered substantial advantages over the original Auxiliary Variable method proposed for the SNF model in Lunagómez et al. (2020). It would be interesting to explore whether other approximations might also confer advantages and develop an appreciation for which methods are best suited for different settings.

More generally, the flexibility of the SNF model offers the potential to address many applied research questions when used to analyse populations of network data. It would be interesting to consider whether the SNF model could be further developed to construct formal hypothesis tests that permit additional model based inferences in this setting. The mathematical challenges to overcome here are non-trivial and present an interesting avenue for future research.

As our desire to analyse more complex data structures increases so do the modelling and computational challenges. Our methodology for incorporating network cycles for statistical modelling and inference of ecological data has opened up an exciting new area within analysis of multiple networks to explore. Accordingly, there is the potential to build on this and address a number of important questions in the field, both theoretical and applied.

## Supplementary information

The supplement to "Bayesian modelling and computation utilising directed cycles in multiple network data" contains additional details about our proposed distance metric, methodology and real data.

**Supplementary Information** The online version contains supplementary material available at <https://doi.org/10.1007/s11222-025-10810-8>.

**Author Contributions** A.M., R.M, S.L. and S.K. contributed to problem conception ; A.M., R.M, S.L. contributed to method development; A.M. contributed to formulating the model and writing the code; A.M. prepared all figures and conducted the simulation study and implementation on real data; A.M. wrote the first draft of the manuscript; All authors reviewed the manuscript.

**Funding** This project was funded by Lancaster University through a Faculty studentship.

**Data Availability** Data can be made available upon request.

**Code Availability** Code is publicly available at: [https://github.com/mantziou/modelling\\_multiple\\_networks\\_using\\_cycles](https://github.com/mantziou/modelling_multiple_networks_using_cycles).

## Declarations

**Conflicts of Interest** To our knowledge there are no competing interests of any form.

**Ethical Standards** Data were collected using behavioural observations in the field, were non-invasive, and did not affect the chance of animal survival.

**Consent to participate** Not applicable.

**Consent for publication** All authors consent for publication of this study.

**Competing interests** The authors declare no competing interests.

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

## References

- Amini, A., Paez, M., Lin, L.: Hierarchical stochastic block model for community detection in multiplex networks. *Bayesian Anal.* **19**(1), 319–345 (2024)
- Arroyo, J., Athreya, A., Cape, J., Chen, G., Priebe, C.E., Vogelstein, J.T.: Inference for multiple heterogeneous networks with a common invariant subspace. *J. Machine Learn. Res.* **22**(142), 1–49 (2021)
- Banks, D., Carley, K.: Metric inference for social networks. *J. Classif.* **11**(1), 121–149 (1994)
- Chen, M.-H., Shao, Q.-M.: On monte carlo methods for estimating ratios of normalizing constants. *Ann. Stat.* **25**(4), 1563–1594 (1997)
- Delmas, E., Besson, M., Brice, M.-H., Burkle, L.A., Dalla Riva, G.V., Fortin, M.-J., Gravel, D., Guimarães, P.R., Jr., Hembry, D.H., Newman, E.A., et al.: Analysing ecological networks of species interactions. *Biol. Rev.* **94**(1), 16–36 (2019)
- Donnat, C., Holmes, S.: Tracking network dynamics: a survey using graph distances. *Annals Appl. Stat.* **12**(2), 971–1012 (2018)
- Durante, D., Dunson, D.B., Vogelstein, J.T.: Nonparametric bayes modeling of populations of networks. *J. Am. Stat. Assoc.* **112**(520), 1516–1530 (2017)
- Ecological Competition. <https://www.encyclopedia.com/environment/energy-government-and-defense-magazines/ecological-competition> (2022)
- Fan, T., Lü, L., Shi, D.: Towards the cycle structures in complex network: a new perspective. *arXiv preprint arXiv:1903.01397* (2019)
- Ginestet, C.E., Li, J., Balachandran, P., Rosenberg, S., Kolaczyk, E.D., et al.: Hypothesis testing for network data in functional neuroimaging. *Annals Appl. Stat.* **11**(2), 725–750 (2017)
- Gollini, I., Murphy, T.B.: Joint modeling of multiple network views. *J. Comput. Graph. Stat.* **25**(1), 246–265 (2016)
- Grether, G.F., Peiman, K.S., Tobias, J.A., Robinson, B.W.: Causes and consequences of behavioral interference between species. *Trends Ecol. Evolut.* **32**(10), 760–772 (2017)
- Han, D., Li, J., Wang, H., Su, X., Hou, J., Gu, Y., Qian, C., Lin, Y., Liu, X., Huang, M., et al.: Circular rna circmtol acts as the sponge of microrna-9 to suppress hepatocellular carcinoma progression. *Hepatology* **66**(4), 1151–1164 (2017)
- Josephs, N., Amini, A.A., Paez, M., Lin, L.: Nested stochastic block model for simultaneously clustering networks and nodes. *arXiv preprint arXiv:2307.09210* (2023)
- Josephs, N., Lin, L., Rosenberg, S., Kolaczyk, E.D.: Bayesian classification, anomaly detection, and survival analysis using network inputs with application to the microbiome. *Annals Appl. Stat.* **17**(1), 199–224 (2023)
- Keith, S.A., Baird, A.H., Hobbs, J.-P.A., Woolsey, E.S., Hoey, A.S., Fadli, N., Sanders, N.J.: Synchronous behavioural shifts in reef fishes linked to mass coral bleaching. *Nat. Clim. Chang.* **8**(11), 986–991 (2018)
- Kinlock, N.L.: Uncovering structural features that underlie coexistence in an invaded woody plant community with interaction networks at multiple life stages. *J. Ecol.* **109**(1), 384–398 (2021)
- Kolaczyk, E.D., Lin, L., Rosenberg, S., Walters, J., Xu, J.: Averages of unlabeled networks: geometric characterization and asymptotic behavior. *Ann. Stat.* **48**(1), 514–538 (2020)
- Koutrouli, M., Karatzas, E., Paez-Espino, D., Pavlopoulos, G.A.: A guide to conquer the biological network era using graph theory. *Front. Bioeng. Biotechnol.* **8**, 34 (2020)
- Laird, R.A., Schamp, B.S.: Competitive intransitivity promotes species coexistence. *Am. Nat.* **168**(2), 182–193 (2006)
- Le, C.M., Levin, K., Levina, E., et al.: Estimating a network from multiple noisy realizations. *Electron. J. Stat.* **12**(2), 4697–4740 (2018)
- Lunagómez, S., Olhede, S.C., Wolfe, P.J.: Modeling network populations via graph distances. *J. Am. Stat. Assoc.* **116**(536), 2023–2040 (2020)
- Mallows, C.L.: Non-null ranking models. i. *Biometrika* **44**(1/2), 114–130 (1957)
- Mantziou, A., Lunagómez, S., Mitra, R.: Bayesian model-based clustering for populations of network data. *Annals Appl. Stat.* **18**(1), 266–302 (2024)
- Mardia, K., Dryden, I.: The complex watson distribution and shape analysis. *J. Royal Stat. Soc.: Ser. B (Stat. Methodol.)* **61**(4), 913–926 (1999)
- Maugis, P.-A.G., Olhede, S.C., Wolfe, P.J.: Topology reveals universal features for network comparison. *arXiv preprint arXiv:1705.05677* (2017)
- Maugis, P.-A., Olhede, S., Priebe, C., Wolfe, P.: Testing for equivalence of network distribution using subgraph counts. *J. Comput. Graph. Stat.* **29**(3), 455–465 (2020)
- Mittelbach, G.G., McGill, B.J.: *Species Interactions in Ecological Networks*, pp. 205–179. Oxford University Press, United Kingdom (2019)
- Mohd, M.H.: Diversity in interaction strength promotes rich dynamical behaviours in a three-species ecological system. *Appl. Math. Comput.* **353**, 243–253 (2019)
- Møller, J., Pettitt, A.N., Reeves, R., Berthelsen, K.K.: An efficient markov chain monte carlo method for distributions with intractable normalising constants. *Biometrika* (2006)
- Newman, M.E.: Estimating network structure from unreliable measurements. *Phys. Rev. E* **98**(6), 062321 (2018)
- Nielsen, A.M., Witten, D.: The multiple random dot product graph model. *arXiv preprint arXiv:1811.12172* (2018)
- Park, J., Haran, M.: Bayesian inference in the presence of intractable normalizing functions. *J. Am. Stat. Assoc.* **113**(523), 1372–1390 (2018)

- Peixoto, T.P.: Reconstructing networks with unknown and heterogeneous errors. *Phys. Rev. X* **8**(4), 041011 (2018)
- Robert, C., Casella, G.: *Monte Carlo Statistical Methods*. Springer, Germany (2013)
- Sizemore, A.E., Giusti, C., Kahn, A., Vettel, J.M., Betzel, R.F., Bassett, D.S.: Cliques and cavities in the human connectome. *J. Comput. Neurosci.* **44**(1), 115–145 (2018)
- Sokhn, N., Baltensperger, R., Bersier, L.-F., Hennebert, J., Ultes-Nitsche, U.: Identification of chordless cycles in ecological networks. In: *International Conference on Complex Sciences*, pp. 316–324 (2012). Springer
- Vitelli, V., Sørensen, Ø., Crispino, M., Frigessi, A., Arjas, E.: Probabilistic preference learning with the mallows rank model. *J. Machine Learn. Res.* **18**(158), 1–49 (2017)
- Wang, S., Arroyo, J., Vogelstein, J.T., Priebe, C.E.: Joint embedding of graphs. *IEEE Trans. Pattern Anal. Machine Intell.* **43**(4), 1324–1336 (2019)
- Young, J.-G., Kirkley, A., Newman, M.E.: Clustering of heterogeneous populations of networks. *Phys. Rev. E* **105**(1), 014312 (2022)

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.