Modeling Network Populations via Graph Distances

Simón Lunagómez, Sofia C. Olhede & Patrick J. Wolfe

Modeling Network Populations via Graph Distances

Simón Lunagómez, Sofia C. Olhede, Patrick J. Wolfe

Department of Mathematics and Statistics, Lancaster University

Institute of Mathematics, École polytechnique fédérale de Lausanne

Departments of Statistics, Computer Science, and Electrical & Computer Engineering, Purdue University.

Corresponding author Simón Lunagómez s.lunagomez@lancaster.ac.uk

Abstract

This article introduces a new class of models for multiple networks. The core idea is to parametrize a distribution on labelled graphs in terms of a Fréchet mean graph (which depends on a user-specified choice of metric or graph distance) and a parameter that controls the concentration of this distribution about its mean. Entropy is the natural parameter for such control, varying from a point mass concentrated on the Fréchet mean itself to a uniform distribution over all graphs on a given vertex set. We provide a hierarchical Bayesian approach for exploiting this construction, along with straightforward strategies for sampling from the resultant posterior distribution. We conclude by demonstrating the efficacy of our approach via simulation studies and two multiple-network data analysis examples: one drawn from systems biology and the other from neuroscience.

Keywords: Hierarchical Bayesian models, Graph metrics, Network variability, Object oriented data, Random graphs, Statistical network analysis.

Introduction

This article introduces a new class of models for data consisting of observations of multiple networks. With advances in measurement technology, these types of data are rapidly becoming prominent in fields such as systems biology and neuroscience, among others. In systems biology, inferences must often be combined on the same gene interaction network, where different inferences correspond to different data sets or to different analysis procedures applied to the same data (Bartlett et al., 2014). In neuroscience, a population of networks encodes the way different regions of the brain interact when individuals perform a given task (Biswal et al., 2010), or characterises a population of individuals suffering from a neurological or psychiatric disorder (Lynall et al., 2010; Nelson et al., 2017).

The developments proposed herein are therefore motivated by the problem of modelling populations of networks. The class of models we propose is based on the idea that distributions on graph space are naturally parameterised in terms of a mean—the Fréchet mean, which is itself a network—and a measure of how concentrated the distribution is about this mean. A benefit of our approach is that the Fréchet mean itself can be interpreted as the representative of a population of networks, relative to a user-specified choice of metric or graph distance. To specify concentration around the Fréchet mean, we use entropy as described below. We then provide general strategies for performing Bayesian inference for these new models, allowing for the modeller to decide which metric is most suitable for the given application at hand.

By multiple networks we mean two or more networks comprising a set of independent observations, which we assume here are defined over the same vertex set. Generalizing this problem to map networks of arbitrary different sizes to a common reference or with scrambled order of nodes is a non-trivial extension, solving a potentially computationally intractable problem (we elaborate

more on this point in the discussion). In for example medical imaging and bioinformatics this assumption is not unreasonable, if admittedly restrictive. A brain connectome example (which we study later in Section 6) drawn from neuroscience is displayed in Fig. 1, with regions of the brain assigned to nodes according to the CC200 atlas, which was proposed by Craddock et al. (2012). Note from Fig. 1 that if we consider each possible pair of observations (first three figures, from left to right), any member of such pair can be seen as a modification of the other member or as a modification from a representative of the population (Fig. 1, on the right). Thus, although one modelling approach would be to treat such networks as realisations from a single random graph model based on global features, such as a stochastic block model, this limits the inferential insights that can be gained from multiple networks as opposed to a single one.

Indeed, the questions arising from multiple network data demand a different perspective:

1. How does one find a summary or representative (at the population level) for multiple observed networks? In other words, what type of structure must the modeller impose on the space of labelled graphs to define a suitable estimand? Without such an estimand (*e.g.*, in the case of block modelling or link prediction) we risk our inference yielding a summary of the population that does not look like any of its elements, and cannot be used in place of them.

2. In the Bayesian setting, if we have multiple networks (such as those in Fig. 1) as historical data, how do we perform prior elicitation without resorting to global assumptions on the network structure? For example, in systems biology it is typical that past inferences regarding a given gene interaction network may provide a very accurate idea about what a newly inferred network might be expected to look like, when obtained using a new measurement technology. This is illustrated in Section 6.1.

We show here that both questions can be answered by first assuming that the observed networks are perturbations of a "typical" network, and then characterising the variability of the data in those terms. Specifically, the Fréchet mean implied by a given metric will parametrize a generative model, under the assumption that the probability of generating a specific network is given by a strictly decreasing function only of its distance from this Fréchet mean.

To construct our models, we borrow ideas from the graphical models and shape theory literatures, where authors have considered the notion of a "typical" non-Euclidean observation, and random perturbations from that observation. Previous work on multiple networks in the statistics literature includes the following: The approaches proposed by Balachandrian et al. (2017) and Chang et al. (2018) for estimating features (subgraph counts and density, respectively) from network data; the model proposed by Gollini and Murphy (2016) (which is an extension of the latent space model proposed by Hoff et al. (2002)) for describing the variability of a homogeneous population of networks; the Bayesian nonparametric approach proposed by Durante et al. (2017) for modelling heterogeneous populations of networks; and the approach for comparing populations of networks via testing by Ginestet et al. (2017). The methodology of the last paper is based on the asymptotic theory for the space of unlabelled networks developed by Kolaczyk et al. (2019), which serves to quantify how concentrated the distribution is around a mean network when formulated in terms of a very specific metric. Kolaczyk et al. (2019) and earlier Feragen et al. (2011) discuss the problem of estimating a mean and the geometry associated to the space of possible values for that estimand, the former for the space of graphs while the latter for the space of trees.

Recently, Nielsen and Witten (2018) proposed a multiple network model based on the random dot product graph model; their approach builds on work by Wang et al. (2017), who proposed a gradient-descent method to compute the simultaneous embedding of a set of graphs. In terms of inference, Nielsen and

Witten (2018) focus on the problem of comparing populations of networks. Tang et al. (2017) focus on the problem of testing for the difference of two populations of networks; the authors assume a random dot product graph model, as in Wang et al. (2017), but computation is done using the bootstrap. We also note the model-based approach for estimating the generating mechanism of multiple networks given by Bhattacharyya and Chatterjee (2018). Finally, some of the ideas developed in this paper have parallels in the literature for modelling measurement error for networks, including recent work by Newman (2018), Peixoto (2018b) and Le et al. (2018).

In a different direction, similarity measures on the local structure of a network have been used to perform prior elicitation on graph space, particularly in the graphical models literature; this idea has been discussed by Mukherjee and Speed (2008) as well as Mitra et al. (2013). Our approach can also be related to work by Tan et al. (2017) and the work by Ni et al. (2018) in the graphical models literature, who propose hierarchical models on graph space. From the shape theory literature, we borrow insight from the work of Mardia and Dryden (1998), which uses the idea of modelling a set of non-Euclidean objects (shapes) in terms of a centroid and parameters that control how concentrated the distribution will be around that centroid.

From a Bayesian point of view, computing a Fréchet mean at the population level is analogous to minimising the posterior expected loss, and becomes the same problem when the loss is a metric. Wade and Ghahramani (2017) exploit this idea, in the context of cluster analysis. In our work, we use entropy in conjunction with the Fréchet mean to define a distribution for non-Euclidean data, and in that sense, our work relates to the methodology developed by Pennec (2006).

Distinct from the literature discussed above, the methodology we propose here achieves different goals: (1) It enables the modeller to characterise the variability of a set of observed networks in terms of a Fréchet mean and a measure of how

concentrated the distribution is around this mean, and to perform Bayesian inference, without resorting to asymptotics; (2) It enables the practitioner interested in network data to perform prior elicitation on graph space by using an observed network as starting point; and (3) It provides tools for incorporating different metrics on graph space into the modelling procedure, enabling the encoding of different assumptions the practitioner may have regarding similarity among graphs. We also are able to discuss a looser notion of a location-scale family for random graph models using this set of technology, taking inspiration from (Fang et al., 1990), we use the functional form of a symmetric multivariate distribution whose both location and scale need to be relaxed to apply. Concrete examples of this notion are provided, along with theoretical results that show that these examples are legitimate. We show how these examples relate to one another.

The remainder of this article is organised as follows. Section 2 first introduces the necessary preliminaries, including metrics on graph space and the Fréchet mean. Section 3 then details the general concepts on which the generative models proposed in this paper are be based on, along some examples. The corresponding strategies for Bayesian modelling and computation are presented in Section 4. Section 5 documents the behaviour of our models via simulation studies, and Section 6 describes fully the fitting of our models to the multiple-network data introduced in Fig. 1 above. Finally, Section 7 discusses briefly the contributions of our approach, placing it in context and outlining limitations as well as future possibilities.

Preliminaries

A simple labelled graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ comprises a set of vertices $\mathcal{V}$ and a set of edges $\mathcal{E} \subset \{ E \subset \mathcal{V} : |E| = 2 \}$. Letting $N = |\mathcal{V}|$, we may represent $\mathcal{G}$ by an $N \times N$ adjacency matrix $A_{\mathcal{G}}$ such that

$$A_{\mathcal{G}}(i, j) = \begin{cases} 1 & \text{if there is an edge between nodes } i \text{ and } j, \\ 0 & \text{otherwise.} \end{cases}$$

The models and methods we propose can all be applied equally to directed graphs (with $A_{\mathcal{G}}(i, j)$ distinct from $A_{\mathcal{G}}(j, i)$ for $i < j$) and those having self-loops ($A_{\mathcal{G}}(i, i) = 1$), as well as more generally any weighted graph such that each $A_{\mathcal{G}}(i, j)$ takes values in some finite, discrete set. We write that a graph $\mathcal{G}_1 = (\mathcal{V}_1, \mathcal{E}_1)$ is a subgraph of $\mathcal{G}_2 = (\mathcal{V}_2, \mathcal{E}_2)$ if $\mathcal{V}_1 \subseteq \mathcal{V}_2$ and $\mathcal{E}_1 \subseteq \mathcal{E}_2$. For a set $\{\mathcal{G}_s\}_{s \in S}$, we denote by $A_{\mathcal{G}_k}(i, j)$ the (i, j)th entry of the adjacency matrix of $\mathcal{G}_k, k \in S$.

For $N \in \mathbb{N}$ and $\mathcal{V} := \{1, 2, \dots, N\}$, define

$$\mathcal{G} := \left\{ \mathcal{G}_{[N]} \right\} := \left\{ \mathcal{G} = (\mathcal{V}, \mathcal{E}) : | \mathcal{V} | = N \right\},$$

so that $\left\{ \mathcal{G}_{[N]} \right\}$ represents the set of all N-node labelled networks of a given type (simple, directed, etc.). If we consider simple graphs, for example, then $\left| \left\{ \mathcal{G}_{[N]} \right\} \right| = 2^{\binom{N}{2}}$. We refer to $\left\{ \mathcal{G}_{[N]} \right\}$ for the simple directed case as a *graph space*. This term tends to be used in this way (rather informally) in the graphical models literature.

Metrics on graph spaces in turn allow for an appropriate definition of network structural similarity (Donnat and Holmes, 2018). We are interested in developing probability models on $\left\{ \mathcal{G}_{[N]} \right\}$ given the choice of a metric $d_G(\cdot, \cdot)$ on $\left\{ \mathcal{G}_{[N]} \right\}$. Two examples of metrics which can be used to formulate the models introduced in Section 3 below are as follows:

1. The *Hamming distance* between two graphs when their adjacency matrices are treated as strings, which is given by the number of entries that disagree. We will use the notation

$$d_H(\mathcal{G}_1, \mathcal{G}_2) = | A_{\mathcal{G}_1} - A_{\mathcal{G}_2} |_H,$$

to denote this distance independently of the type of network under consideration by the modeller (*e.g.*, simple, directed).

2. A *Diffusion distance* based on the graph Laplacian, for example the choice made by <u>Hammond et al.</u> (<u>2013</u>):

$$d_L(\mathcal{G}_1, \mathcal{G}_2; t) = \left(\exp(-tL_{\mathcal{G}_1}) - \exp(-tL_{\mathcal{G}_2}) \right)_F^2, \quad t > 0;$$

where $((_F$ is the Frobenius norm and $L_{\mathcal{G}}$ is the combinatorial Laplacian matrix associated to an undirected graph $\mathcal{G}$:

$$L_{\mathcal{G}}(i, j) = \begin{cases} \sum_{k=1}^{N} A_{\mathcal{G}}(i, k) & \text{if } i = j, \\ -A_{\mathcal{G}}(i, j) & \text{otherwise.} \end{cases}$$

Note that this is referred to by the letter *L* in (<u>Chung</u>, <u>1997</u>), whilst Chung (unlike Hammond et al.) defines the Laplacian to be normalized. The normalized version of the matrix is an operator related to the Laplace-Beltrami operator for objects different than networks via the discretization of a derivative.

Hammond and coauthors (<u>Hammond et al.</u>, <u>2013</u>) argued that the diffusion distance is natural as two graphs are similar if they transmit information in the same way. Generic transmission is by them modelled using heat diffusion on the network. The distance therefore arises as $\exp(-tL_{\mathcal{G}})$ is the kernel associated with (*e.g.*, classical heat) diffusion on a graph $\mathcal{G}$ via the discrete Laplace operator $L_{\mathcal{G}}$. The value of *t* is here the time of diffusion. As $t \to 0$ we should return to whatever initial conditions were specified, and at $t \to \infty$ equal proportions of diffused "stuff" should be at each node.

The value of $d_L(\mathcal{G}_1, \mathcal{G}_2)$ measures the discrepancy after *t* units of time between the diffusion on $\mathcal{G}_1$ versus that on $\mathcal{G}_2$. For our purposes *t* may be regarded as a parameter whose value can be elicited *a priori* using information from the application domain under consideration. As *t* decreases, it becomes harder to distinguish between diffusion patterns (no diffusion has happened yet) and therefore to distinguish between different elements of $\{\mathcal{G}_{[N]}\}$. Finally the graph Laplacian is discussed in detail in (Chung, 1997), and we use the unnormalized version. Hammond et al. (2013) argues that this captures the temporal evolution of the vector representing the diffusion. Thus this metric captured how differently things have flowed up to time *t*.

While the Hamming distance focuses on simple flips of edges into non-edges (changes in very local structure), the diffusion distance is treating the objects functionally (*i.e.* it focuses on changes that may impact the global structure).

One might ask what choice of metric should be made? Donnat and Holmes (2018) provide some guidelines in this choice of metric. The Hamming distance can be interpreted as simple flips of edges (a local modification of the network). The diffusion distance allow information to diffuse on the network, and then lets us compare that diffusion. As Donnat and Holmes (2018) discuss, the Hamming distance assume deletions and additions carry the same weight, even if their structural impact may not be equivalent. The Hamming distance is strongly affected by the sparsity of the graph, as already pointed out by the aforementioned authors. To take into account the sparsity into the metric, the Jaccard distance is used (Donnat and Holmes, 2018). The authors additionally discuss global metrics based on spectral distances. The diffusion distance balances information differently taking the diffusion of information on the network into account. Other balances between global and local can be made (Donnat and Holmes, 2018). Computationcal cost factor into our usage of these two metrics. Later on (see Section 4.2) we will illustrate how a model based on a simple

metric can aid the computation of the posterior for a model based on a more sophisticated metric.

We conclude this section by introducing the Fréchet (1948) mean for use in the context of the metric spaces $(\{\mathcal{G}_{[N]}\}, d_G)$ and associated probability models that we will consider below. Given an arbitrary metric space $(\mathcal{Y}, d)$ and a probability measure on $\mathcal{Y}$, the Fréchet mean provides the notion of an average or measure of central tendency with respect to d. It generalizes the first moment to non-Euclidean settings and has seen wide use in areas such as shape theory.

Definition 2.1 (Fréchet mean). *Let Y be a random element defined on sample space* $\mathcal{Y}$ *and let* $d(\cdot,\cdot)$ *be a metric on* $\mathcal{Y}$ *. The set*

$$\psi^{m} = \arg\inf_{\psi \in \mathcal{Y}} \mathbb{E}_Y[d^2(Y, \psi)] \qquad (1)$$

is called the Fréchet mean set of Y.

We will use the Fréchet mean in conjunction with unimodality to formulate natural and intuitive models on the space $\{\mathcal{G}_{[N]}\}$ of labelled *N*-node networks.

Modelling Approach

In this section, we propose a generative modelling approach for data sets consisting of multiple networks. Our models are parametrized in terms of a unique mode and a univariate measure of dispersion around that mode. The mode in the space of labelled *N*-node networks $\{\mathcal{G}_{[N]}\}$ is itself a network defined on the same vertex set as each individual observation, allowing us to define a suitable estimand to obtain directly a population-level summary of multiple networks.

In analogy to a location–scale family, we provide concepts that enable us to propose probability models on $\{\mathcal{G}_{[N]}\}$ in terms of a central graph (location) and

concentration around that central graph (scale of variation). We use the terms loosely given that we are working in a non-Euclidean setting. In contrast with the location–scale family, which takes the vector space structure for granted, we are constrained by the structure entailed by a metric in $\left\{\mathcal{G}_{[N]}\right\}$ and the fact that the space is finite.

Definition 3.1 (Unimodal network distribution based on location). *Fix a metric d_G on $\left\{\mathcal{G}_{[N]}\right\}$ for $N \in \mathbb{N}$, and consider a family of probability mass functions $\left\{p(\cdot \mid \mathcal{G}^m)\right\}_{\mathcal{G}^m \in \left\{\mathcal{G}_{[N]}\right\}}$ on $\left\{\mathcal{G}_{[N]}\right\}$ such that*

1. *Each $p(\cdot \mid \mathcal{G}^m)$ is unimodal with mode $\mathcal{G}^m \in \left\{\mathcal{G}_{[N]}\right\}$;*
2. *For $\mathcal{G}_1, \mathcal{G}_2 \in \left\{\mathcal{G}_{[N]}\right\}$, we have that $d_G(\mathcal{G}_1, \mathcal{G}^m) > d_G(\mathcal{G}_2, \mathcal{G}^m)$ implies $p(\mathcal{G}_2) > p(\mathcal{G}_1)$, while $d_G(\mathcal{G}_1, \mathcal{G}^m) = d_G(\mathcal{G}_2, \mathcal{G}^m)$ implies $p(\mathcal{G}_2) = p(\mathcal{G}_1)$.*

The most straightforward example is as follows: the Centred Erdös–Rényi Model, which will be introduced later on this section. We shall now need another important concept from information theory, namely that of entropy, see Mézard and Montanari (2009). This is used to measure the uncertainty of a random variable and takes the form of

$$H_{\mathcal{G}} = -\sum_{\mathcal{G} \in \mathfrak{b}} p(\mathcal{G} \mid \mathcal{G}^m) \cdot \log\left(p(\mathcal{G} \mid \mathcal{G}^m)\right). \quad (2)$$

Sometimes $\log(\cdot)$ in the above expression is replaced by $\log_2(\cdot)$. We set $0 \cdot \log(0)$ to equate to zero, as usual.

Building from unimodality we also need to introduce scale, which is our next step.

Definition 3.2 (Unimodal network distribution with location & scale). *Fix a metric* $d_G(\cdot,\cdot)$ *on* $\{\mathcal{G}_{[N]}\}$ *for* $N \in \mathbb{N}$, *nonempty set* $\Gamma \subset \mathbb{R}_+$, *and consider a family* $\{p(\cdot \mid \mathcal{G}^m, \gamma)\}_{\mathcal{G}^m \in \{\mathcal{G}_{[N]}\}, \gamma \in \Gamma}$:

1. *For every fixed scale parameter* $\gamma^* \in \Gamma$, *the family* $\{p(\cdot \mid \mathcal{G}^m, \gamma^*)\}_{\mathcal{G}^m \in \{\mathcal{G}_{[N]}\}}$ *satisfies Definition 3.1 with respect to the metric* d_G.

2. *For every fixed location parameter* $\mathcal{G}^* \in \{\mathcal{G}_{[N]}\}$, *the entropy associated to the family* $\{p(\cdot \mid \mathcal{G}^*, \gamma)\}_{\gamma \in \Gamma}$ *is a strictly monotone function of* $\gamma \in \Gamma$.

For finite, discrete sets such as $\{\mathcal{G}_{[N]}\}$ and associated probability mass function $p(\cdot)$, entropy $-\mathbb{E}\{\log p(\cdot)\}$ provides a convenient characterization akin to variance, ranging from 0 for a point mass to $\log(|\{\mathcal{G}_{[N]}\}|)$ for the uniform distribution on $\{\mathcal{G}_{[N]}\}$. Entropy can thus be used to parametrize a family of discrete distributions on $\{\mathcal{G}_{[N]}\}$ with the same unique mode, in an analogous way to how the scale parameter would parametrize a member of the location–scale family when the location parameter has been specified. The metric provides a ranking of the elements $\{\mathcal{G}_{[N]}\}$ given the mode, the entropy enables the statistician to control the decay of the values of the probability mass function given that ranking. To take the analogy with a Gaussian distribution γ plays the role of $1/\sigma$ for the Gaussian, where σ^2 is the variance. Therefore we expect $\gamma \to 0$ to play the role of $\sigma \to \infty$, or the maximum entropy solution that should be the least peaked. In contrast, $\gamma \to \infty$, we expect to correspond to the minimum entropy solution, and be the most concentrated distribution. Therefore intuitively, we expect the entropy to decay in γ. This allows us to consider the analogy of "peaked" versus "flat" distributions where γ controls the peak.

We now provide two examples for the random graph distribution based on distance and entropy. These examples will be discussed in detail in Sections 4.1 and 4.2, respectively.

We will now introduce a first example of a random graph distribution based on distance and entropy; we call it the Centred Erdös–Rényi Model. The intuition behind this model is that noisy versions of the centroid (which is denoted by $\mathcal{G}^m$) are generated by flipping edges independently at random with probability α. From a modelling perspective, it is sensible to penalize (or constrain) α so it takes values much smaller than the density of $\mathcal{G}^m$; there is little utility for a model where the trend is overwhelmed by noise.

Definition 3.3 (Centred Erdös–Rényi Model). *Given a graph* $\mathcal{G}^m \in \left\{ \mathcal{G}_{[N]} \right\}$ *and* $1/2 > \alpha > 0$ *, consider a model* $p(\cdot | \mathcal{G}^m, \alpha)$ *on* $\left\{ \mathcal{G}_{[N]} \right\}$ *of the form :*

$$\Pr\left(A_{\mathcal{G}}(i,j) = A_{\mathcal{G}^m}(i,j) \right) = 1 - \alpha. \qquad (3)$$

We call this the Centred Erdös–Rényi Model (CER) with mode $\mathcal{G}^m$ *and parameter* α.

Note that $A_{\mathcal{G}}(i,j)$ generating mechanism can also be written as

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

where the $Z(i,j)$'s are iid $\mathrm{Ber}(\alpha)$ for $1 \le i < j \le N$.

This way of describing $A_{\mathcal{G}}(i,j)$'s generating mechanism highlights that edges or flipped to non-edges, or non-edges to edges, with probability α. This clarifies why we expect $\alpha \le \dfrac{1}{2}$, as otherwise we are more likely to flip all edges and not be " centered" at $A_{\mathcal{G}^m}(i,j)$. This condition also will be required in the proofs of Proposition 3.1, which exactly establishes mode etc. Thus, for the Centred

Erdös–Rényi to effectively serve as an error measurement model, the α parameter should be constrained to be be smaller than the edge density parameter of $\mathcal{G}^m$. The condition $1/2 > \alpha > 0$ furthermore ensures that the maximum likelihood estimator (of $\mathcal{G}^m$) will be the graph that minimises the average number of mismatches with respect to the observed networks. For this model, we do not expect the observed graphs to be, on average, of different density than $\mathcal{G}^m$; this is because the error model affects edges and non-edges equally. Observe that if a parametric random graph model is further imposed upon $\mathcal{G}^m$ (*e.g.*, Erdös–Rényi), then this model can be cast into the approach proposed by <u>Newman</u> (<u>2018</u>).

Proposition 3.1. *We let $d_G(\cdot,\cdot)$ denote the Hamming distance on $\{\mathcal{G}_{[N]}\}$. If two graphs $\mathcal{G}_1$ and $\mathcal{G}_2$ are generated from the Centred Erdös–Rényi model with centroid $\mathcal{G}^m \in \{\mathcal{G}_{[N]}\}$ and $0 < \alpha \le 1/2$ then we have that $d_G(\mathcal{G}_1,\mathcal{G}^m) > d_G(\mathcal{G}_2,\mathcal{G}^m)$ implies $p(\mathcal{G}_2) > p(\mathcal{G}_1)$, while $d_G(\mathcal{G}_1,\mathcal{G}^m) = d_G(\mathcal{G}_2,\mathcal{G}^m)$ implies $p(\mathcal{G}_2) = p(\mathcal{G}_1)$. We deduce that $p(\mathcal{G})$ is unimodal, and that the Centred Erdös–Rényi model is a unimodal network distribution based on location and scale.*

As a second example of a unimodal network distribution based on location and scale, we introduce a model motivated by the notion that the similarity with respect to the centroid is made concrete by the choice of $d_G(\cdot,\cdot)$ (*e.g.* the metrics proposed by <u>Zelinka</u> (<u>1975</u>), <u>Hammond et al.</u> (<u>2013</u>), or the ones discussed in <u>Donnat and Holmes</u> (<u>2018</u>)), and covered by our discussion in Section 2 earlier in the paper.

Definition 3.4 (Spherical Network Family). *Given a graph $\mathcal{G}^m \in \{\mathcal{G}_{[N]}\}$, a metric $d_G(\cdot,\cdot)$ on $\{\mathcal{G}_{[N]}\}$, and $\gamma > 0$, we propose:*

$$p(\mathcal{G} \mid \mathcal{G}^m, \gamma) \propto \exp\left\{-\gamma\phi(d_G(\mathcal{G},\mathcal{G}^m))\right\}, \qquad (4)$$

where $\phi(\cdot)$ is a non-negative strictly increasing function such that $\phi(0)=0$. This is the Spherical Network Family with parameters $\mathcal{G}^m$ and γ.

This model is related to the prior introduced by <u>Mitra et al.</u> (<u>2013</u>), which was introduced in the context of graphical modelling. A main difference with respect to their approach is that the Spherical Network Family is aimed to serve as the functional form for both the likelihood and the prior. This model also relates to the similarity measure proposed by <u>Dahl et al.</u> (<u>2017</u>) for random partitions. The normalizing constant for this model is the reciprocal of:

$$Z(\mathcal{G}^m,\gamma) = \sum_{\mathcal{G} \in \{\mathcal{G}_{[N]}\}} \exp\left\{-\gamma\phi(d_G(\mathcal{G},\mathcal{G}^m))\right\}, \qquad (5)$$

here $Z(\mathcal{G}^m,\gamma)$ is known as the *partition function* of $p(\mathcal{G}\mid\mathcal{G}^m,\gamma)$. We observe directly that $Z(\gamma)>0$ as it is a sum of positive terms. Just like the normalizing constant of any probability mass function, as (5) aggregates over $\mathcal{G} \in \{\mathcal{G}_{[N]}\}$, the sum will not be a function directly of $\phi(d_G(\mathcal{G},\mathcal{G}^m))$, only implicitly as the sum will vary depending on the functional form. Therefore $Z(\mathcal{G}^m,\gamma)$ is a positive constant that does not depend on $d_G(\mathcal{G},\mathcal{G}^m)$.

The functional form proposed for the Spherical Network Family (SNF) is inspired by the notion of symmetry of the density discussed in <u>Fang et al.</u> (<u>1990</u>). A random variable X on $\mathcal{X}$ has the symmetry of the density property if its density $p(\cdot\mid\mu,\gamma)$ is of the form

$$p(X\mid\mu,\gamma) = Z^{-1}(\gamma)\cdot\exp\left[-\gamma\phi(d(X,\mu))\right],$$

where $\mu \in \mathcal{X}, \gamma>0, \phi(\cdot)\geq 0$ is a non-decreasing function, $d(\cdot,\cdot)$ is a metric on $\mathcal{X}$.

Proposition 3.2. *The CER is a member of the SNF.*

The above proposition demonstrates that the SNF is not empty. There are some other properties we would like to see, and to be clear on what properties that we desire, let us show that they hold in the following proposition.

Proposition 3.3. *We let $d_G(\cdot,\cdot)$ denote a graph metric on $\{\mathcal{G}_{[N]}\}$. If two graphs $\mathcal{G}_1$ and $\mathcal{G}_2$ are generated from the Spherical Network Family with centroid $\mathcal{G}^m$ and $\gamma \in \mathbb{R}^+$ then we have that $d_G(\mathcal{G}_1,\mathcal{G}^m) > d_G(\mathcal{G}_2,\mathcal{G}^m)$ implies $p(\mathcal{G}_2) > p(\mathcal{G}_1)$, while $d_G(\mathcal{G}_1,\mathcal{G}^m) = d_G(\mathcal{G}_2,\mathcal{G}^m)$ implies $p(\mathcal{G}_2) = p(\mathcal{G}_1)$. As a consequence $p(\mathcal{G})$ is unimodal, and the Spherical Network Family is a unimodal network distribution based on location. In addition, the Spherical Network Family is unimodal network distribution based on location and scale if $\mathbb{V}\mathrm{ar}\left\{\phi\left[d(\mathcal{G},\mathcal{G}^m)\right]\right\} > 0$.*

The next step consists in verifying if the examples we have presented fulfill the condition stated in Proposition 3.3.

Proposition 3.4. *The CER and SNF equipped with the diffusion distance fulfill the condition $\mathbb{V}\mathrm{ar}\left\{\phi\left[d(\mathcal{G},\mathcal{G}^m)\right]\right\} > 0$ when $\phi(\cdot)$ is the identity function.*

The following property of the sample Fréchet mean will provide insight regarding the behavior of the MLE for both models defined above and supports our intuition that the posterior mode will tend to the true value of the Fréchet mean as the sample size increases.

Proposition 3.5. *The sample Fréchet mean in $\{\mathcal{G}_{[N]}\}$ converges to the true Fréchet mean when the later exists and is unique, for $N \in \mathbb{N}$.*

Definition 3.1 is expressed in terms of the mode of the distribution. The following result indicates how the Fréchet mean and mode relate for the Centred Erdös–Rényi Model:

Proposition 3.6. *The mode and Fréchet mean coincide for the Centred Erdös–Rényi Model defined on* $\left\{ \mathcal{G}_{[N]} \right\}, N \in \mathbb{N}$.

For the Spherical Network Family, the Fréchet mean maximises the kernel of the Boltzmann distribution in Eqn (4). This is a direct consequence of Definitions 2.1 and 3.4.

Now, we have enough elements for presenting our model for multivariate network data. Let $N \in \mathbb{N}$ and d_G a metric on $\left\{ \mathcal{G}_{[N]} \right\}$. To describe the variability of a set observations $\left\{ \mathcal{G}_1, \mathcal{G}_2, \dots, \mathcal{G}_n \right\}$ in $\left\{ \mathcal{G}_{[N]} \right\}^n$, we propose a model of the form:

$$p(\mathcal{G}_1, \mathcal{G}_2, \dots, \mathcal{G}_n \mid \mathcal{G}^m, \gamma) = p(\mathcal{G}^m \mid \mathcal{G}^0, \gamma^0) p(\gamma) \prod_{i=1}^{n} p(\mathcal{G}_i \mid \mathcal{G}^m, \gamma), \quad (6)$$

where $p(\cdot \mid \mathcal{G}^m, \gamma)$ is the likelihood, which is given by a unimodal network distribution based on location and scale; $p(\cdot \mid \mathcal{G}^0, \gamma^0)$ is the prior on the mode of the distribution, such prior is also given by a distribution with the same functional form as the likelihood; finally, $p(\gamma)$ is the prior on the entropy of the distribution. One implication of choosing this parametrization is that the inference will be in terms of the population centroid, which is a network by itself. This enables the statistician to perform an operation equivalent to smoothing in graph space.

We propose this model with the aim to represent the variability of a set of observations $\left\{ \mathcal{G}_1, \mathcal{G}_2, \dots, \mathcal{G}_n \right\}$ in $\left\{ \mathcal{G}_{[N]} \right\}^n$ such that, for every pair $\left\{ \mathcal{G}_i, \mathcal{G}_j \right\}$ with $1 \leq i < j \leq n$, $\mathcal{G}_i$ is a small perturbation of $\mathcal{G}_j$ according to d_G. The main assumptions encoded by the model presented in Equation 6 are:

1. The distribution of the observations is assumed to be unimodal *a priori*;

Accepted Manuscript

2. The variability of the observations is characterised in terms of the dispersion around the mode. Such dispersion is defined in terms of d_G a metric on $\{\mathcal{G}_{[N]}\}$;

3. The prior distribution for the mode is assumed to have the same functional form as the likelihood. This implies that it will be unimodal; its mode will be denoted by $\mathcal{G}^0$. We will not assume any structure on $\mathcal{G}^0$, unless we state otherwise.

The first condition is set to guarantee identifiability of the model. The second condition enables the statistician to use the notion of similarity between networks, which can be subject to elicitation, to define variability in the space of graphs, which is, in contrast, very challenging to elicit. The third condition has parallel versions in the functional data analysis literature: we assume a parametric model for the error, with very simple structure, while allowing the trend to be as complex as it needs to be. An alternative approach would be to assume a trend with more defined structure and allow for a richer error structure. We elaborate more on this point in the discussion.

Bayesian Modelling and Computation

In this section, we introduce Bayesian hierarchical models based on the distributions presented in Section 3. For these models, we assume the same functional form for the sampling distribution and for the prior on the Fréchet mean. We also discuss strategies for sampling from the posterior, with emphasis on the case when the normalising constant depends on the Fréchet mean.

Bayesian Inference for the Centred Erdös–Rényi Model

We now discuss a model of the form given in (6) that is inspired by the Centred Erdös–Rényi Model (CER). The intuition behind this model is the following: given a set of observed networks $\{\mathcal{G}_1, \mathcal{G}_2, \dots, \mathcal{G}_n\}$ in $\{\mathcal{G}_{[N]}\}^n$, their variability can be

characterised in terms of the network $\mathcal{G}^m$ that serves as the mode of the distribution and the dispersion around that network. The network $\mathcal{G}^m$ can also be interpreted as the Fréchet mean of $\left\{\mathcal{G}_{[N]}\right\}^n$ implied by the metric and the probability model.

Within this context, the contribution to the likelihood by each observation $\mathcal{G}_i$ is therefore given by:

$$p(\mathcal{G}_i \mid \mathcal{G}^m, \alpha) = \alpha^{d_H(\mathcal{G}_i, \mathcal{G}^m)}(1-\alpha)^{\frac{(N-1)N}{2} - d_H(\mathcal{G}_i, \mathcal{G}^m)}, \quad (7)$$

where $d_H(\cdot,\cdot)$ is the Hamming norm for matrices. Expressions 6 and 7 provide the elements we need to propose the following Bayesian model:

Definition 4.1 (CER/ CER Model). *Let N and n be elements of $\mathbb{N}$, and take*

$0 < \alpha_0 < 1/2$. *The CER/ CER Model is a multivariate network model on $\left\{\mathcal{G}_{[N]}\right\}^n$ of the form*

$$p(\mathcal{G}_1, \mathcal{G}_2, \ldots, \mathcal{G}_n \mid \mathcal{G}^m, \alpha) = \alpha_0^{d_H(\mathcal{G}^m, \mathcal{G}_0)}(1-\alpha_0)^{\frac{(N-1)N}{2} - d_H(\mathcal{G}^m, \mathcal{G}_0)} p(\alpha)$$

$$\times \prod_{i=1}^{n} \alpha^{d_H(\mathcal{G}_i, \mathcal{G}^m)}(1-\alpha)^{\frac{(N-1)N}{2} - d_H(\mathcal{G}_i, \mathcal{G}^m)}, \quad (8)$$

where, the prior $p(\cdot)$ for α is a scaled Beta on $(0, \frac{1}{2})$. Here, $\mathcal{G}_0 \in \left\{\mathcal{G}_{[N]}\right\}$ and $\alpha_0 \in (0,1)$ are the hyperparameters of the model.

We make no assumptions regarding *N* and *n*. Expression (8) is a consequence of the independence of the error, which should be noted. We assume a Beta distribution for *α* is reasonable, since it can be specified in such a way that it is unimodal and favours values close to zero.

Equation 7 proves helpful for understanding the properties of an Erdös–Rényi random graph as an measurement error model. This implies the following properties for the CER/ CER Model:

1. The log-likelihood can be computed using $O(N^2 n)$ operations; this should be kept in mind when performing Bayesian computations, such as MCMC.
2. For α specified, the MLE is the graph $\mathcal{G}^m$ that minimises the average number of mismatches with respect to the observed networks.

The prior for $\mathcal{G}^m$ has $\mathcal{G}_0$ as its mode and its entropy is determined by the Hamming norm and α_0. For the CER/ CER model, the normalizing constant does not depend on either $\mathcal{G}^m$ or α, therefore, samples of the posterior for $(\mathcal{G}^m, \alpha)$ can be obtained via a Metropolis/ Hastings algorithm with a mixture of kernels. To update $A_{\mathcal{G}^m}$, the adjacency matrix associated to $\mathcal{G}^m$, we use the following proposals:

1. Each $A_{\mathcal{G}^m}(i, j)$ changes its value independently to $1 - A_{\mathcal{G}^m}(i, j)$ with probability $0 < \tau < 1$, or stays fixed with probability $1 - \tau$.
2. Each $A_{\mathcal{G}^m}(i, j)$ is sampled independently from a $\mathrm{Ber}\left\{\dfrac{1}{n}\sum_{k=1}^{n} A_{\mathcal{G}_k}(i, j)\right\}$.

To update α we use a mixture of random walks that reflect at 0 and 0.5. For each of these random walks (indexed by k), the proposed value α^* for $\alpha^{(i+1)}$ is given by:

1. $y = \alpha^{(i)} + \zeta^{(i+1)}$, with $\zeta \sim \mathrm{Unif}(-\upsilon_k, \upsilon_k)$, if $0 < y < 0.5$;
2. $-y$, if $y < 0$;
3. $1 - y$, if $y > 0.5$.

The mixture is over $\{\upsilon_1, \upsilon_2, \cdots, \upsilon_K\}$.

Bayesian Inference for Models in the Spherical Network Family

The Spherical Network Family was defined following the intuition that the likelihood should decrease as a function of the distance $d_G(\cdot,\cdot)$ with respect to a graph $\mathcal{G}^m$ that serves as the Fréchet mean. When proposing the functional form, we adopted concepts from the Rotationally Symmetric Family, proposed by Mardia and Dryden (1998). In contrast to the CER/CER model discussed in Section 4.1, more structure is left unspecified and the model presented in this section allows us to specify $d_G(\cdot,\cdot)$. To perform Bayesian inference for $(\mathcal{G}^m, \gamma)$ as described in Definition 3.4, we propose to use a hierarchical model, following the form proposed in Equation 6:

Definition 4.2 (SN/ SN Model). *Let N and n be elements of* $\mathbb{N}$ *and* $d_G(\cdot,\cdot)$ *a metric on* $\left\{ \mathcal{G}_{[N]} \right\}$ *. The SN/ SN Model is a multivariate network model on* $\left\{ \mathcal{G}_{[N]} \right\}^n$ *of the form*

$$
p(\mathcal{G}_1, \mathcal{G}_2, \ldots, \mathcal{G}_n \mid \mathcal{G}^m, \gamma) \propto \exp\left\{ -\gamma_0 \phi(d_G(\mathcal{G}^m, \mathcal{G}_0)) \right\} p(\gamma)
$$

$$
\times \exp\left\{ -\gamma \sum_{i=1}^{n} \phi(d_G(\mathcal{G}_i, \mathcal{G}^m)) \right\}, \tag{9}
$$

where, $p(\cdot)$ *is the prior on γ, which has support on* $\mathbb{R}^+$ *. Here,* $\mathcal{G}_0 \in \left\{ \mathcal{G}_{[N]} \right\}$ *and* $\gamma_0 \in \mathbb{R}^+$ *are the hyperparameters of the model.*

Some features of this model are:

1. The model allows for different specifications of the metric $d_G(\cdot,\cdot)$, which can be chosen with flexibility, for concreteness, *e.g.* distance based on the graph Laplacian, or a metric based on subgraph counts.
2. It is straightforward to set up a Metropolis/ Hastings algorithm to sample from the prior. The Metropolis ratio for updating $\mathcal{G}^{(\cdot)}$ is of the form:

$$H^{(t,t+1)} = \frac{\exp\left\{-\gamma^0 \phi(d_G(\mathcal{G}^{(t+1)}, \mathcal{G}^0))\right\}}{\exp\left\{-\gamma^0 \phi(d_G(\mathcal{G}^{(t)}, \mathcal{G}^0))\right\}} \times \frac{q(\mathcal{G}^{(t)} \mid \mathcal{G}^{(t+1)})}{q(\mathcal{G}^{(t+1)} \mid \mathcal{G}^{(t)})}, \qquad (10)$$

where q is the proposal distribution; here, we are conditioning on the value of γ^0.

3. The argument $\mathcal{G}^m$ that maximises the log of the function:

$$\log\left(\prod_{i=1}^{n} Z(\mathcal{G}^m, \gamma) \times p(\mathcal{G}_i \mid \mathcal{G}^m, \gamma)\right) = -\gamma \sum_{i=1}^{n} \phi(d_G(\mathcal{G}_i, \mathcal{G}^m)), \qquad (11)$$

where γ is specified, coincides with the Fréchet mean of the observed networks when $\phi(x) = x^2$. This follows from applying the definition of a centroid directly.

From a computational perspective, the fact that the normalizing constant for the observations (*i.e.*, the reciprocal of $Z(\cdot)$ in Equation 5) depends on $\mathcal{G}^m$ implies that the Metropolis/ Hastings algorithm cannot be implemented directly for sampling from the posterior of $(\mathcal{G}^m, \gamma)$. For $\mathcal{G}^m$ unspecified, this model falls into the double-intractable constant distributions. Fortunately, sampling from the posterior for the SN/ SN model falls into the setup discussed by Møller et al. (2006). Therefore, the techniques proposed by Møller et al. (2006) and Andrieu and Roberts (2009) can be implemented to sample from the posterior.

The MCMC scheme proposed by Møller et al. (2006) is based on the idea of simulating auxiliary variables $\mathcal{G}_{*,i}$, which are defined on the same sample space as the data $\mathcal{G}_i, 1 \le i \le n$. These variables are sampled so the factors $Z(\mathcal{G}^m, \gamma)^{-n}$ cancel from the Metropolis ratio. We now introduce some additional notation:

$$\vec{\mathcal{G}} = \{\mathcal{G}_1, \ldots, \mathcal{G}_n\} \quad \text{and} \quad \vec{\mathcal{G}}_* = \{\mathcal{G}_{*,1}, \ldots, \mathcal{G}_{*,n}\}.$$

When applied to the SN/ SN Model, the Metropolis ratio for the scheme proposed by Møller et al. (2006) takes the form:

$$H_{(\mathcal{G}^{m,(t+1)},\gamma^{(t+1)}|\mathcal{G}^{m,(t)},\gamma^{(t)})} = \frac{f(\vec{\mathcal{G}}_*^{(t+1)} | \mathcal{G}^{m,(t+1)}, \tilde{\alpha})}{f(\vec{\mathcal{G}}_*^{(t)} | \mathcal{G}^{m,(t)}, \tilde{\alpha})} \times \frac{p(\mathcal{G}^{m,(t+1)} | \mathcal{G}_0, \gamma_0)}{p(\mathcal{G}^{m,(t)} | \mathcal{G}_0, \gamma_0)} \times \frac{p(\vec{\mathcal{G}} | \mathcal{G}^{m,(t+1)}, \gamma^{(t+1)})}{p(\vec{\mathcal{G}} | \mathcal{G}^{m,(t)}, \gamma^{(t)})}$$

$$\times \frac{p(\vec{\mathcal{G}}_*^{(t)} | \mathcal{G}^{m,(t)}, \gamma^{(t)})}{p(\vec{\mathcal{G}}_*^{(t+1)} | \mathcal{G}^{m,(t+1)}, \gamma^{(t+1)})} \times \frac{q(\mathcal{G}^{m,(t)},\gamma^{(t)} | \mathcal{G}^{m,(t+1)},\gamma^{(t+1)})}{q(\mathcal{G}^{m,(t+1)},\gamma^{(t+1)} | \mathcal{G}^{m,(t)},\gamma^{(t)})}, \qquad (12)$$

where the terms of the form:

1. $p(\vec{\mathcal{G}} | \mathcal{G}^{m,(\cdot)}, \gamma^{(\cdot)})$ correspond to the product of kernel of the Boltzmann distribution evaluated at the data $\vec{\mathcal{G}}$, *i.e.*,

$$p(\vec{\mathcal{G}} | \mathcal{G}^{m,(\cdot)}, \gamma^{(\cdot)}) = \exp\left\{ -\gamma^{(\cdot)} \sum_{i=1}^{n} \phi\left[d(\mathcal{G}_i, \mathcal{G}^{m,(\cdot)}) \right] \right\}.$$

   The notation $(\mathcal{G}^{m,(\cdot)}, \gamma^{(\cdot)})$ means that are specified by the state of the chain.

2. $p(\vec{\mathcal{G}}_*^{(\cdot)} | \mathcal{G}^{m,(\cdot)}, \gamma^{(\cdot)})$ correspond to the kernel of the Boltzmann distribution evaluated at the auxiliary variables $\vec{\mathcal{G}}_*$. These auxiliary variables are obtained via a Metropolis-Hastings scheme. This scheme is the same as the one used for sampling from the prior (Equation 10).

3. $p(\mathcal{G}^{m,(\cdot)}, \gamma^{(\cdot)} | \mathcal{G}_0, \gamma_0)$ correspond to the prior for $(\mathcal{G}^m, \gamma)$ evaluated at the state of the chain.

4. $q(\mathcal{G}^{m,(\cdot)},\gamma^{(\cdot)} | \mathcal{G}^{m,(\cdot)}, \gamma^{(\cdot)})$ correspond to the proposal distribution for $(\mathcal{G}^m, \gamma)$. To update $A_{\mathcal{G}^m}$, we use the same hybrid kernel as the one described in Section 4.1. To update γ, the parameter that controls the entropy of the distribution, we use a hybrid kernel formed by a collection of random walks that reflect at 0.

5. $f(\vec{\mathcal{G}}_*^{(\cdot)} | \mathcal{G}^{m,(\cdot)}, \tilde{\alpha})$ correspond to the conditional density of the auxiliary variables. We adopted the probability mass function of the Centred Erdös–

Rényi Model as the conditional density for the auxiliary variables, which are denoted by $(\mathcal{G}^1_*,\dots,\mathcal{G}^n_*)$, *i.e.*,

$$f(\mathcal{G}^1_*,\dots,\mathcal{G}^n_* \mid \mathcal{G}^m,\tilde{\alpha}) = \tilde{\alpha}^{\sum_{i=1}^{n} d_H(\mathcal{G}_i,\mathcal{G}^m)}(1-\tilde{\alpha})^{\frac{(N-1)N}{2}-\sum_{i=1}^{n} d_H(\mathcal{G}_i,\mathcal{G}^m)},$$

as in Section 2 of <u>Møller et al.</u> (<u>2006</u>). Here, $\tilde{\alpha}$ is the posterior mean of the dispersion parameter of a CER/ CER model, which can be estimated as described in Section 4.1. This is the strategy suggested in Equation 7 of <u>Møller et al.</u> (<u>2006</u>).

Some of these terms involve tuning parameters; for instance, $q(\mathcal{G}^{m,(\cdot)},\gamma^{(\cdot)} \mid \mathcal{G}^{m,(\cdot)},\gamma^{(\cdot)})$ requires us to define a mixture of random walks (indexed by *k*), to propose a value γ^* for $\gamma^{(i+1)}$. One way to define such random walks is given by:

1. $y = \gamma^{(i)} + \zeta^{(i+1)}$, with $\zeta \sim \mathrm{Unif}(-\upsilon_k,\upsilon_k)$, if $0 < y$;
2. $-y$, if *y* < 0;

The mixture is over $\{\upsilon_1,\upsilon_2,\cdots,\upsilon_K\}$. Here υ is a tuning parameter. Specifying a prior for *γ* also presents a challenge, since its behaviour will depend drastically on the metric.

In Figure 2 we display the results of a simulation aimed to show the relationship between *γ* and $\mathbb{E}\{\phi[d(\mathcal{G},\mathcal{G}^m)]\}$ for the SN model. These figures serve multiple purposes: (i) to provide information regarding which scales are reasonable for υ, since they provide intuition of how a local change in *γ* would impact a value that is easier to interpret; (ii) to inform where in $\mathbb{R}^+$ the practitioner should allocate most of the mass of the prior for *γ*; (iii) to make informed decisions of how to specify γ^0; (iv) to determine if it is more sensible to keep the prior support for *γ* as $\mathbb{R}^+$, or to constrain it to an interval $(0,\kappa)$. The same applies to the random walks to update *γ*.

Simulation Studies

In this section, we explore the behaviour of the CER/ CER model and the SN/ SN model via simulation studies. We consider that it should be of interest to practitioners to know: i) How precise the inferences become as a function of the number of networks analysed (we will refer to this number as the sample size); ii) To what extent samples from the posterior predictive resemble the data used to obtain the posterior; iii) How sensitive are the inferences with respect to model misspecification. With this in mind, we designed the simulation studies to investigate how the posterior concentrates around the true Fréchet mean as a function of sample size, how regions of high mass of the predictive resemble a neighborhood of the data and robustness.

Concentration of the Posterior as a Function of Sample Size

In this section, we propose simulation experiments to obtain better understanding of how the posterior for $\mathcal{G}^m$ concentrates around its true value as a function of sample size. Ideally, we would like to investigate if the limit

$$\Pr\left\{ d_G(\mathcal{G}, \mathcal{G}^m) > \epsilon \mid \mathcal{G}_1, \mathcal{G}_2, \dots, \mathcal{G}_n \right\} \to 0, \quad (13)$$

holds almost surely as $n \to \infty$, given $\epsilon > 0$, as N is assumed fixed. This is equivalent to asking about the concentration of the posterior, as explained in Section 13.4.1 of Ghosal and van der Vaart (2017). The intuition behind Equation 13 is that, as the sample size n increases, the probability mass of the posterior tends to concentrate on a neighborhood of the true value of the parameter. This statement should be valid for every size of the neighbourhood $\epsilon > 0$. In Equation 13, $\mathcal{G}^m$ is the true value of the mode and $\mathcal{G}$ is a sampled value from the posterior distribution implied by $\{\mathcal{G}_1, \mathcal{G}_2, \dots, \mathcal{G}_n\}$. Equation 13 provides the principle behind the following simulation experiments:

1. Explore how the distance between the point estimate $\mathcal{G}^m$ given by the posterior mode and $\mathcal{G}^m$ behaves as a function of sample size.

2. Investigate how the probability

$$\Pr\left\{ d_G(\mathcal{G}, \mathcal{G}^m) > \epsilon \mid \mathcal{G}_1, \mathcal{G}_2, \dots, \mathcal{G}_n \right\} < \delta, \quad (14)$$

behaves as a function of $n \in \mathbb{N}^+$, here $\epsilon > 0, \delta > 0$ are in turn fixed.

The first simulation provides insight about the speed of convergence of a point estimate (see Fig. 3), while the second simulation investigates how the posterior mass becomes contained in a neighborhood of size ϵ of $\mathcal{G}^m$ as the sample size increases (see Table 2). Here, the size of the neighborhood is controlled by ϵ, and δ serves as a threshold for the amount of posterior mass to be allowed outside the neighborhood.

The simulation regimes are given by:

1. The type of hierarchical model under study (CER/ CER, SN/ SN);
2. The structure imposed on $\mathcal{G}^m$, the centroid of the distribution. These were generated from the Erdös–Rényi model (ER), the Stochastic Block model (SBM), the Small World model (SW), or as a Random Geometric Graph (RGG). The specification of the parameters for these models is displayed in Table 1.

For both the CER/ CER model and the SN/ SN model, we used 250 samples after a burn-in of 100, 000, and a lag of 50. The size of the networks we considered was $N = 50$. The value for γ_0 was specified as 0.01 (for the CER/ CER model, we set $\alpha_0 = 0.01$). A different value of $\mathcal{G}_0$ was obtained for each $\mathcal{G}^m$; it was sampled from $p(\cdot \mid \mathcal{G}^m, \gamma_0)$. This way, we were able to make $\mathcal{G}^m$ exhibit the different types of structure we needed while keeping it as a perturbation of $\mathcal{G}_0$, with concentration given by γ_0 (α_0).

Results from the first and second simulation for the CER/ CER model are summarised in Fig. 3 and Table 2, respectively. These results suggest that, the more homogeneous the adjacency matrix is in terms of inclusion probabilities, the faster the posterior concentrates around the true value. (The Small World and the Stochastic Block models take longer to converge then the Erdös–Rényi and the Random Geometric Graph do.) The Small World model turned out to be especially challenging for our approach, as a consequence of the choice for $\mathcal{G}^0$, which favors a lattice structure. Since the Small World graph is obtained from a re-wiring on a lattice, it takes a larger sample size to disambiguate between the outcomes of the re-wiring process and the perturbation induced by the SN Model.

We compared the performance of our method to the point estimate $\mathcal{G}^m$ we would obtain by computing the majority vote of the data $\{\mathcal{G}_1, \mathcal{G}_2, \dots, \mathcal{G}_n\}$. We used the posterior mode implied by the CER/ CER model to illustrate our method. Results are summarised in Table 3.

Network Prediction

In this section, we investigate the behaviour of our methodology in terms of prediction. We do this according to the following intuition: Given a sample $\{\mathcal{G}_1, \mathcal{G}_2, \dots, \mathcal{G}_n\}$, the posterior predictive distribution should satisfy the criterion that regions with highest posterior density tend to be contained in an open covering of the original sample. To protect ourselves against artifacts due to overfitting, we let the sample used to compute the posterior predictive be distinct from the sample used to compute the open covering. Both from conceptual and computational perspectives, the use of an open covering is valid in this context, since we are working on a metric space of graphs.

We now use this intuition to propose a simulation study. We first generate a sample

$$\{ \mathcal{G}_1, \mathcal{G}_2, \ldots, \mathcal{G}_n, \mathcal{G}_{n+1}, \ldots, \mathcal{G}_{n_t} \}$$

for $(\mathcal{G}^m, \gamma)$ specified, and then we partition this sample into a training set $\{\mathcal{G}_i\}_{i \le n}$ and a test set $\{\mathcal{G}_i\}_{n < i \le n_t}$. Here, $n_t - n$ may be considered a tuning parameter for the simulation, specified by the statistician. Here, the training set will be used to obtain the posterior predictive distribution, while the test set will be used to compute the envelope. In the context of the models we have presented, the assumptions regarding similarity are encoded by the metric $d_G(\cdot, \cdot)$. To make these notions precise, we introduce the tuning parameters $\delta \in (0,1)$ and $\psi_\delta > 0$. Here, ψ_δ is the infimum of $\{\psi : \psi > 0\}$ for which

$$\Pr \left\{ \mathcal{G} \in \bigcup_{k=n+1}^{n} \mathbb{B}(\mathcal{G}_k; \psi) \mid \mathcal{G}_1, \mathcal{G}_2, \ldots, \mathcal{G}_n \right\} \ge 1 - \delta, \quad (15)$$

holds. In Equation 15, $\mathbb{B}(\mathcal{G}_k; \psi)$ denotes the ball with centre $\mathcal{G}_k$ and radius ψ corresponding to $d_G(\cdot, \cdot)$, and $\mathcal{G}$ is a sampled value from the predictive distribution implied by the model and $\{\mathcal{G}_1, \mathcal{G}_2, \ldots, \mathcal{G}_n\}$. The larger ψ_δ is, the less concentrated the posterior predictive distribution will be around the test set. One way to interpret the size of ψ_δ more effectively is by comparing it to quantities for which our intuitions are better informed. We propose comparing it to ρ_δ, the infimum of $\{\rho : \rho > 0\}$ for which

$$\Pr \left\{ \mathcal{G} \in \mathbb{B}(\mathcal{G}^m; \rho) \mid \mathcal{G}^m, \gamma \right\} \ge 1 - \delta,$$

holds, *i.e.*, ρ_δ is the size of the contour set that contains $1 - \delta$ of the probability mass under the specified model.

Implementing this simulation in practice is straightforward: we first compute the distances between each sample from the posterior predictive distribution and the element of $\{\mathcal{G}_{n+1}, \mathcal{G}_{n+2}, \ldots, \mathcal{G}_{n_t}\}$ closest to it. The estimate of ψ_δ is given by the $1 - \delta$

quantile of those distances. Results are summarised in Table 2 for the CER/ CER model and the SN/ SN model.

We used the random graph models and parameter specifications listed in Table 1. We used the same settings for the MCMC (number of samples from the posterior, burn-in, lag) and choices for the hyperparameters ($\mathcal{G}_0$, γ_0 and α_0) as in Section 5.1. The size of the networks was set to $N = 50$.

Results are summarised in Table 4. Here, larger values of ψ_δ indicate that a larger open covering of a sample is needed to mimic regions of the posterior predictive distribution with high probability mass. To be able to compare across regimes, we use the quotient of ψ_δ over ρ_δ, where ρ_δ serves as a quantile. The results in Table 4 suggest that the size of the neighborhood needed to contain the mass of the predictive decays very slowly with respect to the sample size. We also observed that the results were not very sensitive with respect to the generative model for $\mathcal{G}^m$. Recall that ρ_δ is model dependent.

Robustness

In this section, we evaluate the proposed methodology in terms of robustness regarding model misspecification. This is important, since we are making heavily parametric assumptions about the distribution of the deviations with respect to the Frechét mean. We approach this task in two different ways: (i) by using visual diagnostics based on posterior predictive checks (Gelman et al. (1996)), and (ii) by investigating the behaviour of the Bayesian χ^2 (Johnson (2004)) under different scenarios. These methods are further discussed in Appendix B.

The types of misspecification we consider in this simulation study are:

1. Fitting the model when the data was generated by a model based on a different metric on the space of labelled graphs.

2. Fitting the model when the data was generated by a dynamic network model.

For the first type of misspecification, we will fit the SN/ SN model assuming the diffusion distance (Hammond et al., 2013) while the generative model is a CER/ CER model, or vice versa. For the second type of misspecification, we generate data from the dynamic network model implied by making $\mathcal{G}_{k+1}(i, j) \mid \mathcal{G}_{k}(i, j)$ the conditional of a bivariate Bernoulli and then, made all entries of $\mathcal{G}_{k+1}$ conditionally independent given $\mathcal{G}_{k}$, which induces a Markov structure on $\{\mathcal{G}_{1}, \ldots, \mathcal{G}_{n}\}$.

To fit the models, we used the same settings for the MCMC (number of samples from the posterior, burn-in, lag) and choices for the hyperparameters ($\mathcal{G}_{0}$, γ_0 and α_0) as in Sections 5.1 and 5.2. The size of the networks was set to $N = 50$.

Results are summarised in Table 5. In this table, we display the proportion of times where each diagnostic provided evidence for lack of fit over 100 simulated data sets. Both types of diagnostic require us to specify a univariate summary of the of the data. We decided to focus on different quantiles of the degree distribution. The results we obtained suggest that is difficult to assess model misspecification in terms of the center of the degree distribution. It was easier to find evidence of model misspecification, via posterior predictive checks or the Bayesian χ^2, when the focus was on the upper tail of the degree distribution.

Data Analysis

Gene Interaction Data

It has become common practice in systems biology to estimate networks that have either genes or proteins as nodes and where the edges represent, either a potential flow of information (protein signalling) or other evidence of association. Estimating the network is often an intermediate step within a series of inferences and/ or decisions; this is for example the case for the research aimed for the

development of new treatments and vaccines. In this context, having an appropriate characterisation of the variability across different estimated networks can prove key when trying to assess the uncertainty to be associated to the final inferences/ decisions. The variability of the inference of such networks can be due to: i) use of different data bases, ii) use of different technologies to pre-process the data, iii) use of different criteria to decide what constitutes and edge.

An example of a set of networks where the variability is can be attributed to the use of different data bases and/ or technologies is displayed in Figure 4. Here, the nodes stand for the 19 most frequently mutated human cancer genes (the key is provided in Table 6). These genes have a higher-than-expected degree of interconnectivity, this is with respect to sets of genes of similar size selected at random. We consider four types of inferred edges: $N1$ Inferred from expert opinion using curated databases, $N2$ Experimentally determined, $N3$ Obtained via textmining, and $N4$ Obtained via co-expression.

These genes have been widely studied in both the systems biology and cancer research literature. Figure 4 suggests that the set composed by $\{N1, N2, N3\}$ reasonably fulfils the assumptions of our methodology. The edges of $N4$ have a different interpretation, since that graph was obtained via a graphical model. Still, $N4$ can be interpreted as a rough approximation of each element of $\{N1, N2, N3\}$. This data is publicly available from

```
://string-db.org/cgi/network.pl?taskId=PjAoqaYLxdta.
```

Note that nodes 15-19 are isolated. This presents no additional challenge to our methodology since we make no assumptions regarding the connectivity of the observed networks.

We fit the CER/ CER model to the networks $\{N1, N2, N3\}$ and centered the prior for the centroid at $N4$. Results are summarized in Table 7 and Fig. 5. The edge sets corresponding to four networks with highest posterior probability are

displayed in Table 7. The posterior mode is displayed in Fig. 5 (Upper Left), along with summaries for α. We observed that these four networks concentrate more than half of the posterior mass and that the posterior mode concentrates almost 0.25 of the posterior mass. We also observed that nearly 35% of the posterior probability was spread between models (centroids) that were visited by the MCMC only once or twice.

We also fit the SN/ SN model to the data set formed by $\{N1, N2, N3\}$ and centered the prior for the centroid at the minimum spanning tree obtained from assigning random weights to the edges of the graph displayed in Fig. 5 (Upper Left). We centered the prior at this graph instead of using $N4$ because that graph is too far with respect to the data in terms of the graph diffusion distance (Hammond et al. (2013)), for which the creation/ merging of connected components is expensive. In Table 8, we display the three networks with highest posterior probability. We display the posterior mode in Fig. 5 (Upper Right), along with summaries for γ. We observed that these three networks concentrate almost all of the posterior mass and that the posterior mode concentrates more than half of the posterior mass.

The presence of singletons (nodes 15-19) manifests differently in the results, depending on the metric: for the Hamming distance, we observed that the singletons merged to the connected component formed by nodes 1-14 for some of the posterior samples, producing a set of graphs that were visited once or twice by the MCMC, in contrast, when we specified the model in terms the diffusion distance, connected components do not tend to merge or split, which made the set of singletons (nodes 15-19) to remain constant across the MCMC samples.

By fitting both models, we learned that the posterior for the Fréchet mean is sensitive with respect to the metric the model assumes for $\{\mathcal{G}_{[N]}\}$; this becomes evident from comparing Tables 7 and 8 and the two panels at the top of Fig. 5.

The choice of the metric penalises discrepancies between the posterior mode and the Fréchet mean. One way of looking at this, is that, by choosing the metric, the statistician is making decisions regarding which features of the Fréchet mean should be retrieved when computing the posterior. This is a consequence of Proposition 3.5. For this data, we observed an instance of a situation where there are clear differences between choosing $d_G(\cdot, \cdot)$ with input from the practitioner and/ or considerations from the application (SN/ SN model), and choosing the metric based on computational or mathemathical convenience (SER / SER model).

Connectome Data

Connectome data is an instance of measurements of brain activity that are collected, among other purposes: to describe brain structure, to find associations between brain structure and function and to correlate brain structure to covariate information. Among the questions that can be posed given the availability of this type of data, we focus on the following: which is an appropriate representative for either the population or a subpopulation of individuals? One key aspect of this problem consists on making decisions regarding what does it mean for connectomes to be similar. As discussed in Donnat and Holmes (2018), for different metrics in graph space, different representatives and different groupings of the data points may seem appropriate.

We analyzed the dataset discussed in Arroyo et al. (2019) and Zuo et al. (2014). The data consists on 300 instances of connectome data. The connectomes are graphs constructed via diffusion magnetic resonance imaging (dMRI). These measurements were obtained from 30 healthy individuals; 10 measurements were obtained during the curse of a month for each individual. Each of these networks has 200 nodes over the same regions of the brain. The vertices are registered according to the CC200 atlas Craddock et al. (2012). The goal of the analysis performed by Arroyo et al. (2019) was to cluster the graphs according to their community structure (at node level) to see if they could find differences

between individuals. We approach this dataset from a different perspective: we assume the metric based on diffusion and based on that, estimate a representative of the population. We also explore to what extend there is evidence for clusters in the data. We also perform these inferences assuming a Hamming distance.

One of the key assumptions of our methodology is that the data was generated from a unimodal distribution over the space of labelled graphs defined over the same vertex set. The validity of such an assumption depends on the metric. In practice, this assumption can be verified by using a reasoning similar to the one deployed by Donnat and Holmes (2018) when studying the different metrics. We applied multidimensional scaling (MDS) on the data to assess if there is more than one cluster, where each cluster suggests the existence of a different mode. The two-dimensional map for the 300 networks implied by the diffusion distance is displayed in Figure 6. It suggests that modelling the data as unimodal is a reasonable first approximation.

Fitting the CER/ CER model is not a major challenge when analyzing this dataset; the same MCMC scheme as the one used in Section 4.1 can be implemented. For this data set, we used 2,000,000 iterations for burn-in and obtained 5,000 samples with a lag of 1,000. In contrast, fitting the SN/ SN model for a data set is not straightforward. We followed the divide-and-conquer strategy proposed by Wu and Robert (2017). We divided the data into ten subsets of the same size, where each subset preserves the pattern suggested in Figure 6 as much as possible. Each subset is constituted by 30 of the observed networks. We parametrized the model so the posteriors for γ implied by each subset of the data can be transformed into a distribution with roughly same dispersion as the posterior we would obtain by using the whole data set. To compute summaries from the posterior, we proceed as follows:

- For $\mathcal{G}^m$: We let $\mathcal{G}^{m,(i)}$ be the centroid of

$$(\mathcal{G}_1^{m,(i)}, \mathcal{G}_2^{m,(i)}, \ldots, \mathcal{G}_{10}^{m,(i)})$$

with respect to $d(\cdot,\cdot)$. The point estimator for $\mathcal{G}^m$ was obtained by computing the centroid of the posterior modes associated to each subset (as in Section 6.2).

- For γ: since the model is parametrized so γ is on the same scale across the ten subsets. We only need to: i) re-center the all samples with respect to the sample mean of the corresponding subset; ii) re-scale so each posterior has roughly the same dispersion as the full posterior; iii) re-center again using the global sample mean.

For each subset, we ran the MCMC described in Section. We used 500,000 iterations for burn-in and obtained 1,000 samples with a lag of 500.

Results for the Hamming distance are summarized in Figure 7 (Left). Results for the diffusion distance are summarized in Figure 7 (Right). A traceplot of posterior samples for γ corresponding to one of the subsets of the data is displayed in Figure 7. Summaries for γ obtained from combining the samples from the different data sets are also displayed in Figure 7. As a complementary summary, we also show the point estimate for $\mathcal{G}^{m,(i)}$ for one of the subsets of the data in terms of its discrepancies to the point estimate for the whole data set (Figure 8).

Discussion

Network data has caught the imagination of statistical researchers and data analysis practitioners. Despite this interest a number of very fundamental questions lie unresolved in pursuing *multiple* network data analysis. To be able to understand not one network but multiple networks collected simultaneously one has to ask questions like: a) what is the "mean" network (rather than how do we estimate the success-probabilities of an inhomogeneous random graph), and do we want the "mean" itself to be a network? b) what is the degree of variation in

realizations away from that "mean", and how can we make statistical inference in such scenarios? This requires a number of modeling choices, that need to be made for us to make inferences. We in this paper have designed a modular framework that allows us to specify each component, and thus to model.

This modular framework can be compared to the modelling framework of others, such as (Newman, 2018; Le et al., 2018; Chang et al., 2018; Durante et al., 2017; Peixoto, 2018a). In comparison to Durante et al. (2017), for example, we adopt a less flexibly nonparametric approach but allow for our notion of an average or typical network to have complex structure; relative to the approach of (Newman, 2018; Le et al., 2018; Chang et al., 2018; Peixoto, 2018a), by contrast, our parametrisations are more complex while we adopt a similarly simple characterisation of perturbations from the typical network.

The use of the Fréchet mean as a parameter that encodes what the centre of the distribution is supposed to be, as well as the use of the entropy to encode the notion of dispersion, are insights that we borrow exactly from shape theory (Dryden and Mardia, 1998). Even more, the problem of finding a representative for a population of shapes and the problem of modelling the variability of a homogeneous population of shapes are listed as two of the main challenges in that area in Srivastava and Klassen (2016) (Section 1.3). We pose these challenges in the context of network data and offer solutions for the implied inference problems via Bayesian modelling. Some of our theoretical results (Propositions 3 and 4) borrow heavily from shape analysis ideas. From functional data analysis, we adopt the rationale of using a complicated object (a network without a pre-specified structure) to model the trend, while using a simple model to account for the error. The trade-off between the complexity of the trend and the complexity of the error distribution has been widely studied in the functional data analysis literature; a similar tension will arise in our context. In this setting the mean function is often left mainly unspecified (or even just restricted to a form of regularity such as Besov regularity), but the noise is not permitted much

structure. The noise or perturbation from that network we chose to be very simple, normally just uncorrelated white noise. This could be construed as the Goldilocks principle at work, where things are made complex, but not too complex, rather just right in their complexity to capture realistic features. This remains a topic for exploration and/ or future developments. One interesting challenge that arises in the context of network data is that there is a lot to be learned regarding which metric in graph space should be adopted for a given problem. This is an interesting contrast to functional data analysis, since in that context, practitioners are more familiar with the idea of pairing a specific metric to a given application (such as the ℓ^p norm for signal processing).

There are also inevitably limits of resolvability to this problem, linked to being able to resolve the blocks of the stochastic block model (Hajek et al., 2017). Here we see identifiability starts to depend on the number of nodes, and the observed number of networks, as well as the level of variability of each individual network. Our study of small world networks, show that if the number of observed networks are sufficiently few, then the regularizing effect of the prior can indeed be too strong.

In Section 1 we mentioned that it would be challenging to extend the proposed methods to a setting where the vertex set is allowed to vary in an unconstrained manner. However, if the vertex set varies so it is always a subset of a maximal finite collection of vertices, then the methods proposed in this paper are still valid (provided an appropriate metric is provided) and the the theory results will hold. In terms of computation, an MCMC based on a saturated model approach (as the one proposed in Section 5 of Brooks et al. (2003)) can be used to obtain samples from the posterior. The key aspect here is that, for this setting, we still have a finite discrete space endowed with a metric, which is the core assumption for our method.

In contrast to the methodology proposed by Durante et al. (2017), which focuses on clustering, our methodology in turn is designed for providing summaries that are easy to interpret in the context of replication and on prior elicitation, in addition, our methodology makes explicit what the estimand for a central network is, instead of just providing an estimator with no obvious estimand associated to it. The main advantage of our method with respect to approaches that use the idea of a Fréchet mean as a centre, but derive the uncertainty around that centre via asymptotics (Ginestet et al. (2017)) are: i) that our method enables the statistician to propagate uncertainty to subsequent inferences, since we are able to sample from a posterior, and ii) our method is not constrained to use of a single metric, in contrast to Ginestet et al. (2017), which relies on a specific metric to derive the asymptotic results they need. In a broad sense, this last point also applies to the approach proposed by Durante et al. (2017), since their MCMC scheme relies heavily on the metric induced by a random dot product model to take advantage of conjugacy.

The proposed methodology can help in the development of informative priors for graphical models. The example discussed in Section 6.1 suggests how to proceed: i) obtain the posterior mode from previous/ similar studies; ii) apply the proposed methodology with a metric that can be related to a measure of similarity in distribution space (such as the Kullback-Leibler divergence); and iii) use the posterior produced this way as the prior for the data associated to the graphical model we want to infer.

Future work includes: i) to develop methodology that enables the use of mixture distributions at the level of the centroid network. There are two possibilities for achieving this: to specify the number of elements in the mixture (hierarchical model approach) or to leave the number of elements unspecified (the Bayesian nonparametric approach); ii) to extend the current methodology to allow for missing data and/or partial observation of the network due to sampling. This would raise interesting challenges, since in our approach the network is treated

as the observational unit; iii) to constrain the structure of the centroid by using a parametric model (such as Peixoto (2018b); Newman (2018)), or to impose specific constrains on graph features of the centroid. Such an extension demands a formulation in terms of hierarchical models. By constraining the possible values for the centroid, we should be able to propose richer models for the error distribution.

From our perspective, to get a better understanding of the trade-offs between imposing structure on the centroid versus imposing structure for the error distribution is a promising area for future research. It is not straightforward to anticipate which combinations of assumptions for the centroid and the error distribution will lead to useful models, since, both, the use of metrics on a graph space and the use of random graphs as error models have not been explored from a statistician's perspective.

References

C. Andrieu and G. O. Roberts. The pseudo-marginal approach for efficient Monte Carlo computations. *Annals of Statistics*, 37(2):697–725, 2009.

J. Arroyo, J. Athreya, A.and Cape, G. Chen, C. Priebe, and J. T. Vogelstein. Inference for multiple heterogeneous networks with common invariant subspace. arXiv preprint arXiv:1906.10026, 2019.

P. Balachandrian, E. D. Kolaczyk, and W. D. Viles. On the propagation of low-rate measurement error to subgraph counts in large networks. *Journal of Machine Learning and Research*, 18(61): 1–33, 2017.

T. E. Bartlett, S. C. Olhede, and A. Zaikin. A DNA methylation network interaction measure and detection of network oncomarkers. *Plos One*, 9(1), January 2014.

S. Bhattacharyya and S. Chatterjee. Spectral clustering for multiple sparse networks: I. *arXiv preprint arXiv:1805.10594*, 2018.

B. B. Biswal, M. Menness, and X.-N. Zuo. Toward discovering science of human brain function. *Proceedings of the National Academy of Sciences of the United States of America*, 107(10): 4734–4739, 2010.

S.P. Brooks, P. Giudici, and G.O. Roberts. Efficient construction of reversible jump markov chain monte carlo proposal distributions. *Journal of the Royal Statistical Society: Series B*, 65(1): 3–39, 2003.

J. Chang, E. D. Kolaczyk, and Q. Yao. Estimation of edge density in noisy networks. arXiv preprint arXiv:1803.02488v1, 2018.

F.K.R. Chung. *Spectral Graph Theory*. American Mathematical Society, Providence, Rhode Island, 1997.

R. C. Craddock, G. James, P. E. Holtzheimer, X. P. Hu, and H. S. Mayberg. A whole brain fmri atlasgenerated via spatially constrained spectral clustering. *Human Brain Mapping*, 33(8): 1914–1928, 2012.

D.B. Dahl, R. Day, and J.W. Tsai. Random partition distribution indexed by pairwise information. *Journal of the American Statistical Association*, 112(518):721–732, 2017.

C. Donnat and S. Holmes. Tracking network ddynamic: a survey using graph distances. *The Annals of Applied Statistics*, 12(2):971–1012, 2018.

I. L. Dryden and K. V. Mardia. *Statistical Shape Analysis*. John Wiley & Sons, 1998.

D. Durante, D. B. Dunson, and J. T. Vogelstein. Nonparametric bayes modeling of populations of networks. *Journal of the American Statistical Association*, 112(520):1516–1530, 2017.

K. T. Fang, S. Kotz, and Ng K. W. *Symmetric Multivariate and Related Distributions*. Chapman and Hall Ltd, 1990.

A. Feragen, S. Hauberg, M. Nielsen, and F. Lauze. Means in space of tree-like shapes. In *2011 Internation Conference on Computer Vision*, 2011.

M. Fréchet. Les éléments aléatories de nature quelconque dans un espace distancié. *Annales de l'Institut Henri Poincaré*, 10:103–111, 1948.

A. Gelman, X.-L. Meng, and H. Stern. Posterior predictive assessment of model fitness via realized discrepancies. *Statistica Sinica*, 6:733–807, 1996.

S. Ghosal and A. van der Vaart. *Fundamentals of Nonparametric Bayesian Inference*. Number 44 in Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2017.

C.E. Ginestet, J. Li, P. Balachandrian, S. Rosenberg, and E. D. Kolaczyk. Hypothesis testing for network data in functional neuroimaging. *Annals of Applied Statistics*, 11(2):725–750, 2017.

I. Gollini and T. B. Murphy. Joint modeling of multiple network views. *Journal of Computational and Graphical Statistics*, 25(1):246–265, 2016.

Bruce Hajek, Yihong Wu, and Jiaming Xu. Information limits for recovering a hidden community. *IEEE Transactions on Information Theory*, 63(8):4729–4745, 2017.

D. K. Hammond, Y. Gur, and C. R. Johnson. Graph diffusion distance: A difference measure for weighted graphs based on the graph Laplacian exponential kernel. In *Global Conference on Signal and Information Processing (GlobalSIP), 2013 IEEE*, 2013.

P. D. Hoff, A. E. Raftery, and M. S. Handcock. Latent space approaches to social network analysis. *Journal of the American Statistical Association*, 97(460):1090–1098, 2002.

V. E. Johnson. A Bayesian chi^2 test for goodness-of-fit. *Annals of Statistics*, 32(6):29–46, 2004.

E. D. Kolaczyk, L. Lin, S. Rosenberg, and J. Walters. Averages of unlabeled networks: Geometric characterizchara and asymptoasy behavior. *Annals of Statistics*, 2019. (to appear).

C.M. Le, K. Levin, and E. Levina. Estimating a network from multiple noisy realizations. *Electronic Journal of Statistics*, 12(2):4697–4740, 2018.

M.-E. Lynall, D. S. Bassett, and R. Kerwin. Functional connectivity and brain networks in Schizophrenia. *Journal of Neuroscience*, 30(28):9477–9487, 2010.

K. V. Mardia and I. L. Dryden. The complex Watson distribution and shape analysis. Technical report, Department of Statistics, University of Leeds, 1998.

M. Mézard and A. Montanari. *Information, Physics, and Computation*. Oxford University Press, 2009.

R. Mitra, P. Müller, S. Liang, L. Yue, and Y. Ji. A Bayesian graphical model for ChiP-Seq data on Histone modifications. *Journal of the American Statistical Association*, 108(501):69–80, 2013.

J. Møller, A. N. Pettitt, R. Reeves, and K. K. Berthelsen. An efficient Markov chain Monte Carlo method for distributions with intractable normalising constants. *Biometrika*, 93(2):451–458, 2006.

S. Mukherjee and T. P. Speed. Network inference using informative priors. *Proceedings of the National Academy of Sciences of the United States of America*, 105(38):14313–14318, 2008.

B.G. Nelson, J. Bassett, D. S.and CamBass, E.T. Bullmore, and K.O. Lim. Comparison of large-scale human brain functional and anatomical networks in schizophrenia. *NeuroImage:Clinical*, 15: 439–448, 2017.

M. E. J. Newman. Network reconstruction and error estimation with noisy network data. arXiv preprint arXiv:1703.07376v2, 2018.

Y. Ni, P. Müller, Y. Zhu, and Y. Ji. Heterogeneous reciprocal graphical models. *Biometrics*, 74: 606–615, 2018.

A. Nielsen and D. Witten. The multiple random dot product graph model. arXiv preprint arXiv:1811.12172v1, 2018.

T. P. Peixoto. Reconstructing networks with unknown and heterogeneous errors. arXiv preprint arXiv:1806.07956, 2018a.

T. P. Peixoto. Network structure from rich but noisy data. *Physical Review X*, 8(4), 2018b.

X. Pennec. Intrinsic statistics on riemannian manifolds: Basic tools for geometric measureents. *Journal of Mathematical Imaging and Vision*, 25:127–154, 2006.

A. Srivastava and E.C. Klassen. *Functional and Shape Data Analysis*. Springer-Verlag, 2016.

L. Tan, A. Jasra, M. De Iorio, and T. M. D. Ebbels. Bayesian inference for multiple gaussian graphical models with application to metabolic association networks. *Annals of Applied Statistics*, 11(4): 2222–2251, 2017.

M. Tang, A. Athreya, D.L. Sussman, V. Lyzinski, Y. Park, and C. Priebe. A semiparametric two-sample hypothesis testing problem for random graphs. *Journal of Computational and Graphical Statistics*, 26(2):344–354, 2017.

S. Wade and Z. Ghahramani. Bayesian cluster analysis: Point estimation and credible balls. *Bayesian Analysis*, 2017.

S. Wang, J. T. Vogelstein, and C.E. Priebe. Joint embedding of graphs. arXiv preprint arXiv:1703.03862, 2017.

C. Wu and C. Robert. Average of recentered parallel mcmc for big data. arXiv preprint arXiv:1706.04780v2, 2017.

B. Zelinka. On a certain distance between isomorphism classes of graphs. *Casopis Pest. Mat.*, 100: 371–373, 1975.

X.-N. Zuo, J. S. Anderson, P. Bellec, R. M. Birn, B. B. Biswal, J. Blautzik, J. C. Breitner, R. L. Buckner, V. D. Calhoun, F. X. Castellanos, A. Chen, B. Chen, J. Chen, X. Chen, S. J. Colcombe, W. Courtney, R. C. Craddock, A. Di Martino, H.-M. Dong, X. Fu, Q. Gong, K. J. Gorgolewski, Y. Han, Y. He, Y. He, E. Ho, A. Holmes, X.-H. Hou, J. Huckins, T. Jiang, Y. Jiang, W. Kelley, C. Kelly, M. King, S. M. LaConte, J. E. Lainhart, X. Lei, K. Li, H.-J.and Li, K. Li, Q. Lin, D. Liu, J. Liu, X. Liu, Y. Liu, G. Lu, J. Lu, B. Luna, J. Luo, D. Lurie, Y. Mao, D. S. Margulies, A. R. Mayer, T. Meindl, M. E. Meyerand, Nielsen Nan, W., O'Connor J. A., D. D., Paulsen, V. Prabhakaran, Z. Qi, J. Qiu, C. Shao, Z. Shehzad, W. Tang, A. Villringer, H. Wang, K. Wang, D. Wei, G.-X. Wei, X.-C. Weng, X. Wu, T. Xu, N. Yang, Z. Yang, Y.-F. Zang, L. Zhang, Q. Zhang, Z. Zhang, Z. Zhang, K. Zhao, Z. Zhen, Y. Zhou, X.-T. Zhu, and M. P. Milham. An open science resource for establishing reliability and reproducibility in functional connectomics. *Scientific Data*, 2014.

Proofs

Proof of Proposition 3.1

Let $\mathcal{G}_1$ and $\mathcal{G}_2$ in $\{\mathcal{G}_{[N]}\}$. Here $\mathcal{G}^m \in \{\mathcal{G}_{[N]}\}$ is fixed and $d_H(\cdot,\cdot)$ denotes the Hamming distance. Let $N_e = \binom{N}{2}$ be the total number of edges possible in the graph. It follows:

$$
\frac{p\left(\mathcal{G}_1 \mid \mathcal{G}^m, \alpha\right)}{p\left(\mathcal{G}_2 \mid \mathcal{G}^m, \alpha\right)} = \frac{\alpha^{d_H(\mathcal{G}_1,\mathcal{G}^m)}(1-\alpha)^{N_e - d_H(\mathcal{G}_1,\mathcal{G}^m)}}{\alpha^{d_H(\mathcal{G}_2,\mathcal{G}^m)}(1-\alpha)^{N_e - d_H(\mathcal{G}_2,\mathcal{G}^m)}}
$$

$$
= \left(\frac{\alpha}{1-\alpha}\right)^{d_H(\mathcal{G}_1,\mathcal{G}^m) - d_H(\mathcal{G}_2,\mathcal{G}^m)},
$$

$$
= \left(\frac{1-\alpha}{\alpha}\right)^{d_H(\mathcal{G}_2,\mathcal{G}^m) - d_H(\mathcal{G}_1,\mathcal{G}^m)}.
$$

If $0.5 > \alpha > 0$, then $\dfrac{1-\alpha}{\alpha} > 1$. Under this condition, $d_H(\mathcal{G}_2,\mathcal{G}^m) > d_H(\mathcal{G}_1,\mathcal{G}^m)$ if and only if:

$$
\frac{p\left(\mathcal{G}_1 \mid \mathcal{G}^m, \alpha\right)}{p\left(\mathcal{G}_2 \mid \mathcal{G}^m, \alpha\right)} > 1. \qquad (16)
$$

Since $d_H(\cdot,\cdot)$ is a metric, this reasoning implies that the distribution $p(\cdot \mid \mathcal{G}^m,)$ is unimodal. The proof for the case $d_H(\mathcal{G}_1,\mathcal{G}^m) = d_H(\mathcal{G}_2,\mathcal{G}^m)$ follows *mutatis mutandis*.

Our second task is to show that the Centred Erdös–Rényi graph defined on $\{\mathcal{G}_{[N]}\}$ fulfills Definition 3.2. For this, we need to investigate how the entropy $H_{CER(\mathcal{G}^m,\alpha)}$ of the distribution relates to α. Remember that, if X and Y are independent random variables, then the entropy of their joint distribution $H_{X,Y}$ and the entropy of the individual variables (H_X and H_Y) relate as follows:

$$
H_{X,Y} = H_X + H_Y.
$$

as explained in Following <u>Mézard and Montanari</u> (<u>2009</u>), Section 1.2. The CER model with parameters $(\mathcal{G}^m, \alpha)$ can be represented as a random vector of size $N_e = \dfrac{(N-1)N}{2}$, where entries iid $\mathrm{Ber}(\alpha)$. Therefore, the entropy of the CER with parameters $(\mathcal{G}^m, \alpha)$ is given by:

$$H_{\mathrm{CER}(\mathcal{G}^m, \alpha)} = -N_e \times \left[(1-\alpha)\log(1-\alpha) + \alpha\log(\alpha)\right].$$

Since

$$\frac{\partial}{\partial\gamma} H_{\mathrm{CER}(\mathcal{G}^m, \alpha)} = -N_e \times \left[\log\left(\frac{\alpha}{1-\alpha}\right)\right],$$

we conclude that the entropy of the distribution is a strictly increasing function of α in $[0, 0.5]$, with $H_{\mathrm{CER}(\mathcal{G}^m, \alpha)} = 0$ and such that the maximum entropy is reached at $\alpha = 0.5$, for which all elements in $\left\{\mathcal{G}_{[N]}\right\}$ are assigned equal mass (See Example 1.6 from <u>Mézard and Montanari</u> (<u>2009</u>) (2009)). This computation implies that part (2) of Definition 3.2 is fulfilled. Part (1) of Definition 3.2 is fulfilled since the proof that CER model satisfies Definition 3.1 was carried out for α specified and no property of $\mathcal{G}^m$ was invoked or constrain on it was imposed.

Proof of Proposition 3.2

Proof. Let us start from the PMF of the graph. Let N be the number of nodes in the graph, and N_e the total number of possible edges. Let n_j be the number of switches of $\mathcal{G}_j$ away from $\mathcal{G}^m$. We can then write

$$p\left(\mathcal{G}_j \mid \mathcal{G}^m, \alpha\right) = \alpha^{n_j}(1-\alpha)^{N_e - n_j}$$

$$= \exp\left\{n_j \log(\alpha) + (N_e - n_j)\log(1-\alpha)\right\} \qquad (17)$$

$$= \exp\{N_e \log(1-\alpha)\}\exp\left\{n_j \log\left(\frac{\alpha}{1-\alpha}\right)\right\}.$$

We note directly that

$$n_j = (A_{\mathcal{G}_j} - A_{\mathcal{G}^m}) \big|_H = d_H(\mathcal{G}_j, \mathcal{G}^m).$$

We note that

$$p\left(\mathcal{G}_j \mid \mathcal{G}^m, \alpha\right) = \exp\{N_e \log(1-\alpha)\} \exp\left\{ d_H(\mathcal{G}_j, \mathcal{G}^m) \log\left(\frac{\alpha}{1-\alpha}\right) \right\}.$$

We therefore see that we have

$$\phi(x) = x,$$

with $\gamma = \log(\dfrac{1-\alpha}{\alpha})$, or $e^\gamma = \dfrac{1-\alpha}{\alpha}$, or $1/\alpha = e^\gamma + 1$. Thus

$$\alpha = \frac{1}{1+e^\gamma} \Rightarrow 1-\alpha = \frac{e^\gamma}{1+e^\gamma} \Rightarrow \log(1-\alpha) = \gamma - \log(1+e^\gamma).$$

and so

$$
\begin{aligned}
p\left(\mathcal{G}_j \mid \mathcal{G}^m, \gamma\right) &= \exp\{N_e \log(1-\alpha)\} \exp(-\gamma\phi(d_H(\mathcal{G}_j, \mathcal{G}^m))) \\
&= \exp\{N_e[\gamma - \log(1+e^\gamma)]\} \exp(-\gamma\phi(d_H(\mathcal{G}_j, \mathcal{G}^m))) \qquad (18) \\
&= \frac{\exp\{N_e\gamma\}}{(1+e^\gamma)^{N_e}} \exp(-\gamma\phi(d_H(\mathcal{G}_j, \mathcal{G}^m)))
\end{aligned}
$$

Thus the CER is a member of the spherical network family (Definition 3.4). We find with the abbreviations Centered Erdős–Renyi (CER), Spherical Network Family (SNF), Unimodal network Distribution based on location (UDL), as well as unimodal network distribution based on location and scale (UDLS), that there is a natural nestedness

$$\mathrm{CER} \in \mathrm{SNF} \subset \mathrm{UDLS} \subset \mathrm{UDL}. \qquad (19)$$

From these, the only inclusion that requires some qualification is $\mathrm{SNF} \subset \mathrm{UDLS}$. This point is addressed in Proposition 3.3. □

Proof of Proposition 3.3

Proof. First, we will prove that Spherical Network Family defined on $\left\{ \mathcal{G}_{[N]} \right\}$ fulfills Definition 3.1. We start with property (2). Let $(\mathcal{G}^m, \gamma)$ be pre-specified. Let $\mathcal{G}_1$ and $\mathcal{G}_2$ be such that:

$$d_G(\mathcal{G}_1, \mathcal{G}^m) > d_G(\mathcal{G}_2, \mathcal{G}^m), \qquad (20)$$

or $\mathcal{G}_1$ is further from $\mathcal{G}^m$ than $\mathcal{G}_2$.

The Spherical Network Family is a Boltzmann distribution on the space of graphs. Boltzmann (or Gibbs) distributions take the form of

$$p(\mathcal{G} \mid \mathcal{G}^m, \gamma) = Z^{-1}(\mathcal{G}^m, \gamma) \exp\left\{ -\gamma \phi(d_G(\mathcal{G}, \mathcal{G}^m)) \right\}, \qquad (21)$$

where $Z^{-1}(\mathcal{G}^m, \gamma)$ is a normalizing constant ($Z(\mathcal{G}^m, \gamma)$ is the partition function) and $\gamma > 0$. We note that maximum entropy is approached as $\gamma \to 0$ (when the distribution becomes uniform) and the distribution degenerates to a point mass at $\mathcal{G}^m$ as $\gamma \to \infty$. This means that, in the limit cases, the model has the desired behaviour.

The relationship of (20) occurs if and only if

$$\phi(d_G(\mathcal{G}_1, \mathcal{G}^m)) > \phi(d_G(\mathcal{G}_2, \mathcal{G}^m)$$
$$\Leftrightarrow -\gamma \phi(d_G(\mathcal{G}_1, \mathcal{G}^m)) < -\gamma \phi(d_G(\mathcal{G}_2, \mathcal{G}^m))$$
$$\Leftrightarrow \exp\left\{ -\gamma \phi(d_G(\mathcal{G}_1, \mathcal{G}^m)) \right\} < \exp\left\{ -\gamma \phi(d_G(\mathcal{G}_2, \mathcal{G}^m)) \right\}$$
$$\Leftrightarrow p(\mathcal{G}_1 \mid \mathcal{G}^m, \gamma) < p(\mathcal{G}_2 \mid \mathcal{G}^m, \gamma),$$

as $\exp\{\cdot\}$ is a strictly increasing function and the constants of proportionality cancel. The argument for the equality case follows *mutatis mutandis*.

We now proceed to prove Definition 3.1(1) holds: Let $\mathcal{G}_1 \neq \mathcal{G}^m$, then $d_G(\mathcal{G}_1, \mathcal{G}^m) > 0$ since $d_G(\cdot, \cdot)$ is a metric. It follows that $d_G(\mathcal{G}_1, \mathcal{G}^m) > d_G(\mathcal{G}^m, \mathcal{G}^m)$, since $d_G(\mathcal{G}^m, \mathcal{G}^m) = 0$. We conclude that:

$$p(\mathcal{G}_1 \mid \mathcal{G}^m, \gamma) < p(\mathcal{G}^m \mid \mathcal{G}^m, \gamma),$$

which shows that $\mathcal{G}^m$ is a mode and that the mode is unique.

Second, we will prove that Spherical Network Family defined on $\{\mathcal{G}_{[N]}\}$ fulfills Definition 3.2, that is, that the parameter γ controls the entropy of the distribution. This is to achieve the analogy of a Gaussian distribution. The parameter γ is indexing the family of distributions that we study. Any member of the family is characterised by its entropy. We want the indexing to be such that if γ increases, the entropy decreases, and the distribution becomes better concentrated.

To be able to understand the spherical network family, we shall study the so-called Boltzmann or Gibbs distributions. Boltzmann distributions are common in statistical mechanics, and further discussed in (Mézard and Montanari, 2009), Section 2.2. The SNF falls in this class, as is directly apparent from Definition 3.4.

Our next objective is to investigate the entropy of the spherical network family. We first introduce some notation. Let:

$$K(\mathcal{G}, \mathcal{G}^m, \gamma) = \exp\left\{-\gamma\phi\left[d(\mathcal{G}, \mathcal{G}^m)\right]\right\};$$

this function is decreasing in γ. Let:

$$Z(\mathcal{G}^m, \gamma) = \sum_{\mathcal{G} \in \{\mathcal{G}_{[N]}\}} \exp\left\{-\gamma\phi\left[d(\mathcal{G}, \mathcal{G}^m)\right]\right\}.$$

denote the partition function. It follows that both $Z(\mathcal{G}^m, \gamma)$ and $\log\{Z(\mathcal{G}^m, \gamma)\}$ are decreasing in γ, as $Z(\mathcal{G}^m, \gamma)$ is the sum of decreasing functions.

The entropy of a member of the spherical network family is given by:

$$
\begin{aligned}
H_{\text{SNF}} &= -\sum_{\mathcal{G} \in \{\mathcal{G}_{[N]}\}} p(\mathcal{G} \mid \mathcal{G}^m, \gamma) \log\{p(\mathcal{G} \mid \mathcal{G}^m, \gamma)\} \\
&= -\sum_{\mathcal{G} \in \{\mathcal{G}_{[N]}\}} p(\mathcal{G} \mid \mathcal{G}^m, \gamma) \log\left\{\frac{1}{Z(\mathcal{G}^m, \gamma)} \exp\left\{-\gamma\phi\left[d(\mathcal{G}, \mathcal{G}^m)\right]\right\}\right\} \\
&= -\sum_{\mathcal{G} \in \{\mathcal{G}_{[N]}\}} p(\mathcal{G} \mid \mathcal{G}^m, \gamma) \left[\log\{K(\mathcal{G}, \mathcal{G}^m, \gamma)\} - \log\{Z(\mathcal{G}^m, \gamma)\}\right] \quad (22) \\
&= \log\{Z(\mathcal{G}^m, \gamma)\} + \gamma\sum_{\mathcal{G} \in \{\mathcal{G}_{[N]}\}} \phi\left[d(\mathcal{G}, \mathcal{G}^m)\right] p(\mathcal{G} \mid \mathcal{G}^m, \gamma) \\
&= \log\{Z(\mathcal{G}^m, \gamma)\} + \gamma \times \mathbb{E}\left\{\phi\left[d(\mathcal{G}, \mathcal{G}^m)\right]\right\}.
\end{aligned}
$$

The next task is to determine under which conditions, the entropy of the spherical network family is decreasing in γ. We introduce some additional notation. Let $F(\gamma)$ denote the free energy:

$$
F(\gamma) = -\frac{1}{\gamma}\log\{Z(\mathcal{G}^m, \gamma)\}.
$$

The following identity is a standard result for the Boltzmann distribution Mézard and Montanari (2009), p. 25-29.

$$
\mathbb{E}\left\{\phi\left[d(\mathcal{G}, \mathcal{G}^m)\right]\right\} = \frac{\partial}{\partial\gamma}[\gamma F(\gamma)],
$$

which implies

$$
\mathbb{E}\left\{\phi\left[d(\mathcal{G}, \mathcal{G}^m)\right]\right\} = -\frac{\partial}{\partial\gamma}\log\{Z(\mathcal{G}^m, \gamma)\}. \quad (23)
$$

We now compute the derivative of the entropy of the spherical network family with respect to γ. From Equation 22, we have:

$$\frac{\partial}{\partial \gamma} H_{\text{SNF}} = \frac{\partial}{\partial \gamma} \log \left\{ Z(\mathcal{G}^m, \gamma) \right\} + \frac{\partial}{\partial \gamma} \left[\gamma \times \mathbb{E} \left\{ \phi \left[d(\mathcal{G}, \mathcal{G}^m) \right] \right\} \right]$$

$$= - \mathbb{E} \left\{ \phi \left[d(\mathcal{G}, \mathcal{G}^m) \right] \right\} + \mathbb{E} \left\{ \phi \left[d(\mathcal{G}, \mathcal{G}^m) \right] \right\} + \gamma \times \frac{\partial}{\partial \gamma} \mathbb{E} \left\{ \phi \left[d(\mathcal{G}, \mathcal{G}^m) \right] \right\} \quad (24)$$

$$= \gamma \times \frac{\partial}{\partial \gamma} \mathbb{E} \left\{ \phi \left[d(\mathcal{G}, \mathcal{G}^m) \right] \right\}, \quad (25)$$

where the equality in Expression 24 follows from applying Equation 23. By definition,

$$\mathbb{E} \left\{ \phi \left[d(\mathcal{G}, \mathcal{G}^m) \right] \right\} = \frac{1}{Z(\mathcal{G}^m, \gamma)} \sum_{\mathcal{G} \in \left\{ \mathcal{G}_{[N]} \right\}} \phi \left[d(\mathcal{G}, \mathcal{G}^m) \right] \exp \left\{ -\gamma \phi \left[d(\mathcal{G}, \mathcal{G}^m) \right] \right\}. \quad (26)$$

From Equation 26, we obtain that $\dfrac{\partial \mathbb{E} \left\{ \phi \left[d(\mathcal{G}, \mathcal{G}^m) \right] \right\}}{\partial \gamma}$ is equal to

$$- \frac{\frac{\partial Z(\mathcal{G}^m, \gamma)}{\partial \gamma}}{Z(\mathcal{G}^m, \gamma)^2} \sum_{\mathcal{G} \in \left\{ \mathcal{G}_{[N]} \right\}} \phi \left[d(\mathcal{G}, \mathcal{G}^m) \right] K(\mathcal{G}, \mathcal{G}^m, \gamma) + \frac{1}{Z(\mathcal{G}^m, \gamma)} \sum_{\mathcal{G} \in \left\{ \mathcal{G}_{[N]} \right\}} (-1) \phi^2 \left[d(\mathcal{G}, \mathcal{G}^m) \right] K(\mathcal{G}, \mathcal{G}^m, \gamma).$$

Therefore

$$\frac{\partial \mathbb{E} \left\{ \phi \left[d(\mathcal{G}, \mathcal{G}^m) \right] \right\}}{\partial \gamma} = \frac{-1}{Z(\mathcal{G}^m, \gamma)} \sum_{\mathcal{G} \in \left\{ \mathcal{G}_{[N]} \right\}} \left(\phi \left[d(\mathcal{G}, \mathcal{G}^m) \right] + \frac{\frac{\partial Z(\mathcal{G}^m, \gamma)}{\partial \gamma}}{Z(\mathcal{G}^m, \gamma)} \right) \phi \left[d(\mathcal{G}, \mathcal{G}^m) \right] K(\mathcal{G}, \mathcal{G}^m, \gamma)$$

$$= - \mathbb{E} \left\{ \left(\phi \left[d(\mathcal{G}, \mathcal{G}^m) \right] - \mathbb{E} \left\{ \phi \left[d(\mathcal{G}, \mathcal{G}^m) \right] \right\} \right) \phi \left[d(\mathcal{G}, \mathcal{G}^m) \right] \right\}$$

$$(27)$$

$$= - \mathbb{V}\text{ar} \left\{ \phi \left[d(\mathcal{G}, \mathcal{G}^m) \right] \right\} < 0. \quad (28)$$

Here, Equation 27 follows from Equation 23. By applying Equation 28 to Equation 25, we obtain

$$\frac{\partial}{\partial \gamma} H_{\text{SNF}} = -\gamma \times \mathbb{V}\text{ar}\left\{\phi\left[d(\mathcal{G}, \mathcal{G}^m)\right]\right\} < 0. \qquad (29)$$

It follows that the spherical network family is parametrized in terms of $\mathcal{G}^m$, which is the mode of the distribution, and γ, which is a monotone function of the entropy as long as $p(\cdot \mid \mathcal{G}^m, \gamma)$ is not a point mass. We also have that each of these parameters can be specified without constrains imposed by the other, therefore Definition 3.2 is satisfied. □

For further understanding of the behaviour of the entropy as a function of α we refer to the Proof of Proposition 3.1, earlier in this appendix.

Proof of Proposition 3.4

Proof. Let $N \geq 2$. This means that $A_{\mathcal{G}^m}$ has at least one entry in its upper-triangular section, outside of the diagonal; this also means that there is at least one graph $\mathcal{G}_1 \in$, such that $\mathcal{G}_1 \neq \mathcal{G}^m$. For both, the Hamming distance and the diffusion distance, we have:

$$d(\mathcal{G}^m, \mathcal{G}^m) = 0 \qquad \text{and} \qquad 0 < d(\mathcal{G}^m, \mathcal{G}_1) < \infty,$$

and

$$p(\mathcal{G}^m \mid \mathcal{G}^m, \gamma) > p(\mathcal{G}_1 \mid \mathcal{G}^m, \gamma) > 0,$$

for both models. Remember that, for the CER, γ is a function of $\alpha \in (0,1)$. It follows, that $\mathbb{E}\left\{d(\mathcal{G}, \mathcal{G}^m)\right\} > 0$ and therefore $\mathbb{V}\text{ar}\left\{d(\mathcal{G}, \mathcal{G}^m)\right\} > 0$. □

Proof of Proposition 3.5

Proof. For _n_ observations $\{y_1, \dots y_n\}$ from $\{\mathcal{G}_{[N]}\}$ The sample Frechét mean is given by:

$$\hat{\psi}_n = \arg\min_{\psi \in \mathcal{Y}} \frac{1}{n} \sum_{i=1}^{n} d(y_i, \psi)^2; \qquad (30)$$

see Eqn 1. By sampling elements of $\{\mathcal{G}_{[N]}\}$ via a distribution with full support, each individual expectation in Eqn 1 is the limit of the corresponding sample mean. These sample means are part of the computation in Equation 30.

Let ψ be an element of $\{\mathcal{G}_{[N]}\}$ such that $\mathbb{E}(d^2(Y, \psi))$ is finite. We only need to considers those $\psi \in \{\mathcal{G}_{[N]}\}$ for which the expectation is finite, since, both, the Frechét mean and the sample Frechét mean are obtained by computing the minimum. Note that, if $\mathbb{E}(d^2(Y, \psi)) = \infty$ for all $\psi \in \{\mathcal{G}_{[N]}\}$, the assumption of a unique Frechét mean would not be fullfilled. Let _n_ be the number of observed networks, note that, as $n \to \infty$:

$$\frac{1}{n} \sum_{i=1}^{n} d(y_i, \psi)^2 \to \mathbb{E}(d^2(Y, \psi)), \quad \text{a.s. for all} \quad \psi \in \{\mathcal{G}_{[N]}\} \quad \text{such that} \quad \mathbb{E}(d^2(Y, \psi)) < \infty,$$

by the Strong Law of Large Numbers, as $d^2(Y, \psi)$ is a scalar. Here, ψ is fixed and the y_i's are random. This argument tells us that each individual expectation is the limit of the corresponding sample mean, the next part of the argument is to prove that all the expectations in Equation Eqn 1 can be estimated simultaneously with enough accuracy (encoded by ϵ), so the minimization entailed by 30 can be carried out without errors with high probability.

For every $\psi \in \{\mathcal{G}_{[N]}\}$ and all $\epsilon > 0$ and $\delta \in (0,1)$, there exists $N(\epsilon, \delta) \in \mathbb{N}$ such that:

$$\Pr\left\{ \left\| \frac{1}{n} \sum_{i=1}^{n} d(y_i, \psi)^2 - \mathbb{E}(d^2(Y, \psi)) \right\| < \epsilon \right\} > 1 - \delta,$$

for all $n > N(\epsilon, \delta)$. Since $\left\{ \mathcal{G}_{[N]} \right\}$ is finite, this is true for

$$\epsilon < \frac{1}{2} \min_{(\psi_1, \psi_2) \in \left\{ \mathcal{G}_{[N]} \right\} \times \left\{ \mathcal{G}_{[N]} \right\}} \left\| \mathbb{E}(d^2(Y, \psi_1)) - \mathbb{E}(d^2(Y, \psi_2)) \right\|,$$

and $0 < \delta < 1$ pre-specified, for all $\psi \in \left\{ \mathcal{G}_{[N]} \right\}$. This is, we can make the noise of the sampled means smaller than any pairwise difference of the expectations, for ψ_1, ψ_2 in $\left\{ \mathcal{G}_{[N]} \right\}$. Since $\left\{ \mathcal{G}_{[N]} \right\}$ is finite, we can make $N(\epsilon, \delta)$ constant with respect to $\psi \in \left\{ \mathcal{G}_{[N]} \right\}$ by taking the maximum. This means, that for all $n > N(\epsilon, \delta)$,

$$\Pr\left\{ \hat{\psi}_n = \psi^m \right\} = 1 - \delta,$$

where ψ^m is the Fréchet mean, which we assumed to be unique. This is because, for $n \geq N(\epsilon, \delta)$, each expectation can be approximated with enough precision that the optimization can be carried out without error with high probability. □

Proof of Proposition 3.6

Proof. The objective is to prove that, for the CER with parameters $(\mathcal{G}^m, \gamma)$, the mode $\mathcal{G}^m$ coincides with the Fréchet mean. We divide the proof into two parts: for the first part, we provide a condition for when the inner product between a $s-$ dimensional vector *a* with non-negative entries and a pmf *w* is minimized, where the optimization is taken over all permutations of indices for the entries of *a, i.e.,* $\{1, 2, \dots, s\}$; for the second part, we prove that taking the expectation of the distances with respect to a graph $\mathcal{G}_k$ for the CER $(\mathcal{G}^m, \alpha)$ is an example of the setup described in the first part, even more, the permutations of the vector of distances involved in computing the expectation, correspond to different choices for $\mathcal{G}_k$. We conclude by proving that the minimum of the expectation is attained at $\mathcal{G}^m$.

Part 1

Let a and w be vectors with s entries, in addition, let w be such that $w_i > 0$ and $\sum w_i = 1$ and a be such that $a_i \geq 0$. Let σ be a permutation of $\{1, 2, \ldots, s\}$ such that $w_{\sigma[i]} \geq w_{\sigma[j]}$ for every pair $\{i, j\}$ with $i < j$. Let τ be a permutation of $\{1, 2, \ldots, s\}$ such that $a_{\tau[i]} \leq a_{\tau[j]}$ for every pair $\{i, j\}$ with $i < j$. Therefore, τ fullfils:

$$\tau = \arg\min_{\varphi \in \mathrm{Sym}(s)} \sum a_{\varphi[i]} \times w_{\sigma[i]}$$

where $\mathrm{Sym}(s)$ denotes the set of permutations over s indices.

We proof the last statement by induction:

for $s = 2$

: Let us start with the case $a_{\tau[1]} < a_{\tau[2]}$ and $w_{\sigma[1]} > w_{\sigma[2]}$. Since the entries of w are nonnegative and add to 1, it follows that $w_{\sigma[1]} > \dfrac{1}{2} > w_{\sigma[2]}$. Therefore

$$a_{\tau[1]} \times w_{\sigma[1]} + a_{\tau[2]} \times w_{\sigma[2]} \qquad (31)$$

is closer to $a_{\tau[1]}$ than it is to $a_{\tau[2]}$. If one permutes the indices of a_{τ} to obtain a new vector $a_{\tau'}$, then, it follows that

$$a_{\tau'[1]} \times w_{\sigma[1]} + a_{\tau'[2]} \times w_{\sigma[2]} \qquad (32)$$

is closer to $a_{\tau[2]}$ than it is to $a_{\tau[1]}$; one way to visualize this argument is to note that Expressions 31 and 32 correspond to convex linear combinations of two non-negative numbers, namely $(a_{\tau[1]}, a_{\tau[2]})$ and the statements about closeness correspond to the size of the weights $(w_{\sigma[1]}, w_{\sigma[2]})$. Since $a_{\tau[1]} < a_{\tau[2]}$, the condition is fulfilled. For the cases where either $a_{\tau[1]} = a_{\tau[2]}$ or $w_{\sigma[1]} = w_{\sigma[2]}$, it is trivial to show that the condition is fulfilled.

for $s = k$

: Let us assume that the result holds for $s = k$.

for

$s = k + 1$: We consider two cases, which are defined in terms of the existence of fixed points of τ.

Case 1. At least one entry in a_τ remains fixed.

WLOG we can assume that the entry of a_τ that remained invariant is the $(k + 1) -$ th. We can re-normalize the first k entries of w_σ by making

$$w'_{\sigma[i]} = \frac{w_{\sigma[i]}}{1 - w_{\sigma[k+1]}}.$$

Since $w'_{\sigma[i]} \geq w'_{\sigma[j]}$ for all $1 \leq i < j \leq k$, we can apply the hypothesis of induction to the first k entries of $a_\tau, a_{\tau'}$ and w_σ to obtain

$$\sum_{i=1}^{k} a_{\tau[i]} \times w'_{\sigma[i]} \leq \sum_{i=1}^{k} a_{\tau'[i]} \times w'_{\sigma[i]}. \qquad (33)$$

Equation 33 is valid for any permutation τ' that leaves the $k + 1$ entry unchanged when compared to τ. Therefore

$$\sum_{i=1}^{k} a_{\tau[i]} \times w_{\sigma[i]} \leq \sum_{i=1}^{k} a_{\tau'[i]} \times w_{\sigma[i]}$$

since we only need to multiply w'_σ by a positive constant, namely $1 - w_{\sigma[k+1]}$. We assumed $a_{\tau[k+1]} = a_{\tau'[k+1]}$, it follows that

$$\sum_{i=1}^{k+1} a_{\tau[i]} \times w_{\sigma[i]} \leq \sum_{i=1}^{k+1} a_{\tau'[i]} \times w_{\sigma[i]}.$$

We conclude that τ minimises $\sum a_{\tau[i]} \times w_{\sigma[i]}$ for all the permutations that leave at least one entry unchanged with respect to τ.

Case 2. We now consider the case where no entry of a_τ was left invariant by a new indexing τ'. Let a_{τ^*} be the vector that results from permuting two entries of $a_{\tau'}$ so the $(k+1)-$th entry of a_{τ^*} coincides with the $(k+1)-$th entry of a_τ. By applying an argument analogous to the one made for $k = 2$, we obtain:

$$\sum_{i=1}^{k+1} a_{\tau^*[i]} \times w_{\sigma[i]} \leq \sum_{i=1}^{k+1} a_{\tau'[i]} \times w_{\sigma[i]}.$$

Now, since at least the $(k+1)-$th entry of a_{τ^*} coincides with the $(k+1)-$th entry of a_τ, we have

$$\sum_{i=1}^{k+1} a_{\tau[i]} \times w_{\sigma[i]} \leq \sum_{i=1}^{k+1} a_{\tau^*[i]} \times w_{\sigma[i]},$$

therefore the conclusion is valid for this case also.

Part 2

We start by proving that, given $N \in \mathbb{N}$, the number of graphs in $\{\mathcal{G}_{[N]}\}$ such that $d_H(\mathcal{G}, \mathcal{G}^m) = h$, where $h \in \{1, 2, \dots, N_e\}$, is constant with respect to $\mathcal{G}^m$. Having $d_H(\mathcal{G}, \mathcal{G}^m) = h$ implies that h entries of the adjacency matrix of $\mathcal{G}^m$ were modified. This is equivalent from choosing h entries from the upper triangular of the adjacency matrix of $\mathcal{G}^m$. Since the graphs are labelled, the number of graphs such that $d_H(\mathcal{G}, \mathcal{G}^m) = h$ is $\binom{N_e}{h}$, which is constant with respect to $\mathcal{G}^m$. The same argument can be made for every value of $h \in \{1, 2, \dots, N_e\}$.

The expectation $\mathbb{E}\{d_H^2(\mathcal{G}, \mathcal{G}_k)\}$ for a CER with parameters $(\mathcal{G}^m, \alpha)$ can be computed as follows:

$$\mathbb{E}\left\{d_H^2(\mathcal{G}, \mathcal{G}_k)\right\} = \sum_{i \in I} d_H^2(\mathcal{G}_i, \mathcal{G}_k) \times p(\mathcal{G}_i \mid \mathcal{G}^m, \alpha),$$

where I is an indexing for $\left\{\mathcal{G}_{[N]}\right\}$ such that $p(\mathcal{G}_i \mid \mathcal{G}^m, \alpha) \geq p(\mathcal{G}_j \mid \mathcal{G}^m, \alpha)$ for $i < j$. Given the fact that the number of graphs in $\left\{\mathcal{G}_{[N]}\right\}$ that fulfill $d_H(\mathcal{G}, \mathcal{G}^m) = h$ is constant with respect to $\mathcal{G}^m$ for every $h \in \{1, 2, \dots, N_e\}$, the vector

$$\left(d_H^2(\mathcal{G}_1, \mathcal{G}_k), d_H^2(\mathcal{G}_2, \mathcal{G}_k), \dots, d_H^2(\mathcal{G}_{|\{\mathcal{G}_{[N]}\}|}, \mathcal{G}_k)\right)$$

is obtained from permuting the entries from

$$\left(d_H^2(\mathcal{G}_1, \mathcal{G}^m), d_H^2(\mathcal{G}_2, \mathcal{G}^m), \dots, d_H^2(\mathcal{G}_{|\{\mathcal{G}_{[N]}\}|}, \mathcal{G}^m)\right),$$

Now:

$$\mathbb{E}\left\{d_H^2(\mathcal{G}, \mathcal{G}^m)\right\} = \sum_{i \in I} d_H^2(\mathcal{G}_i, \mathcal{G}^m) \times p(\mathcal{G}_i \mid \mathcal{G}^m, \alpha). \qquad (34)$$

The vectors on the right side of Equation 34 and the indexing I fulfill the assumptions of Part 1 (see proof of Proposition 3.1). This implies that $\mathcal{G}^m$ is the Fréchet mean for the CER with parameters $(\mathcal{G}^m, \alpha)$. $\square$

Diagnostics for Bayesian Models

Posterior predictive checks (Gelman et al. (1996)) are based on following the intuition: if the model assumptions are reasonable, the observed value of a statistic should, with low probability, be extreme with respect to the predictive distribution for that statistic. One way to translate this intuition to our context is the following: Let $\eta^{(0)}$ be a one-dimensional summary of the observed networks $\{\mathcal{G}_1, \mathcal{G}_2, \dots, \mathcal{G}_n\}$ (e.g., the average diameter, the average number of communities). Obtain K Monte Carlo data sets $\{\mathcal{G}_1^{(i)}, \mathcal{G}_2^{(i)}, \dots, \mathcal{G}_n^{(i)}\}$, $1 \leq i \leq K$, from the posterior predictive distribution. For each of these data sets, we compute $\eta^{(i)}$, a realisation

of the predictive distribution of the one-dimensional summary, $1 \leq i \leq K$. If $\eta^{(0)}$ is extreme with respect to the Monte Carlo predictive distribution implied by $\{\eta^{(i)}\}_{1 \leq i \leq K}$, then we can regard this as evidence for lack of fit.

The Bayesian χ^2 was proposed by Johnson (2004) and it is based on the following rationale: Each sample from the posterior $(\mathcal{G}^{m,(i)}, \gamma^{(i)})$ entails the distribution of a univariate summary Y, *i.e.*,

$$(\mathcal{G}^{m,(i)}, \gamma^{(i)}) \rightarrow F_Y(\cdot \mid \mathcal{G}^{m,(i)}, \gamma^{(i)}).$$

In the context of multivariate modelling for networks, such summary is a descriptive statistic that can be computed efficiently, *e.g.*, the mean of the degree distribution. Let $(y_1, y_2, \dots, y_n)$ be the observed values for this summary, with $\mathcal{G}_s \rightarrow y_s$. Given a partition

$$0 = a_0 < a_1 < \dots < a_{D-1} < a_D = 1,$$

of the interval $[0,1)$, we can compute the counts

$$C_k^{(i)} = \sum_{j=1}^{n} \mathbb{I}_{[a_{k-1}, a_k)}(F_Y(y_j \mid \mathcal{G}^{m,(i)}, \gamma^{(i)})),$$

for $k \in \{1, 2, \dots, D\}$. Let $p_k = a_k - a_{k-1}$, then

$$R^B(\mathcal{G}^{m,(i)}, \gamma^{(i)}) = \sum_{k=1}^{D} \left(\frac{C_k^{(i)} - np_k}{\sqrt{np_k}} \right)^2$$

measures the discrepancy between the observed and expected counts for the bins $[a_{k-1}, a_k), k \in \{1, 2, \dots, D\}$. Goodness-of-fit is assessed via q/ q plots of $R^B(\cdot)$ with respect to a χ^2 with $D - 1$ degrees of freedom.
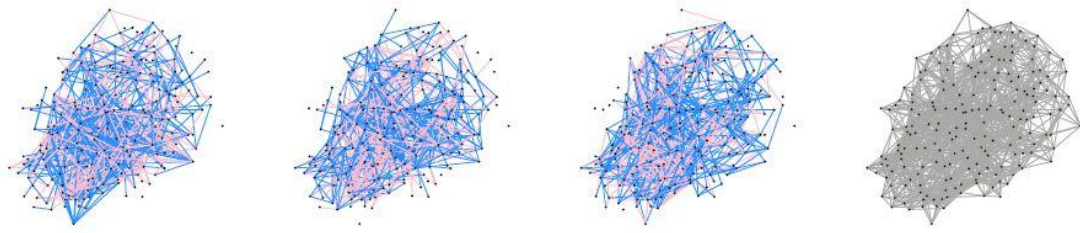
Fig. 1 Example of multiple network data in the context of neuroscience, with each node representing a region of the brain (see Zuo et al. (2014) and Arroyo et al. (2019)). The networks are defined over the same set of 200 nodes. First three figures (from Left to Right): Discrepancies of three observed brain networks with respect to the point estimate of the Frechét mean. Edges only present in the observed network are colored in blue, while edges present in the point estimate but not in the data point are colored in pink. Right: Posterior mode estimate $\mathcal{G}^{FM}$ of the Frechét mean of 300 brain networks using a metric on graph space based on diffusion.

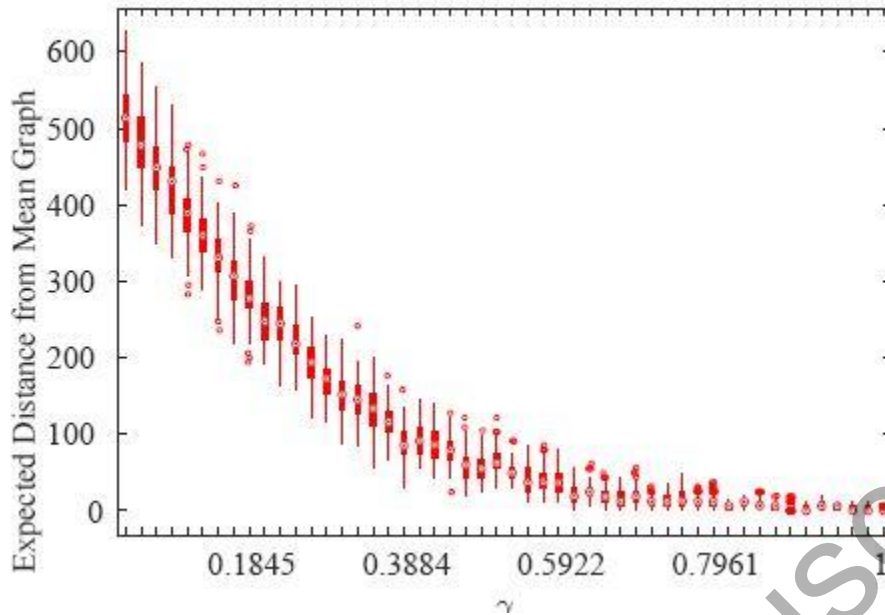**Fig. 2** Distribution of the distance (summarised as a boxplot) to $\mathcal{G}^m$ for the SN model as a function of γ. Here N = 19, t = 1 and $\mathcal{G}^m$ was specified as the network displayed in Figure 4. The figure shows that, for $\gamma > 1$ increments on that parameter do not have a perceptible effect on the distribution. This plot serves to inform the scales that are relevant for defining random walk proposals for γ.
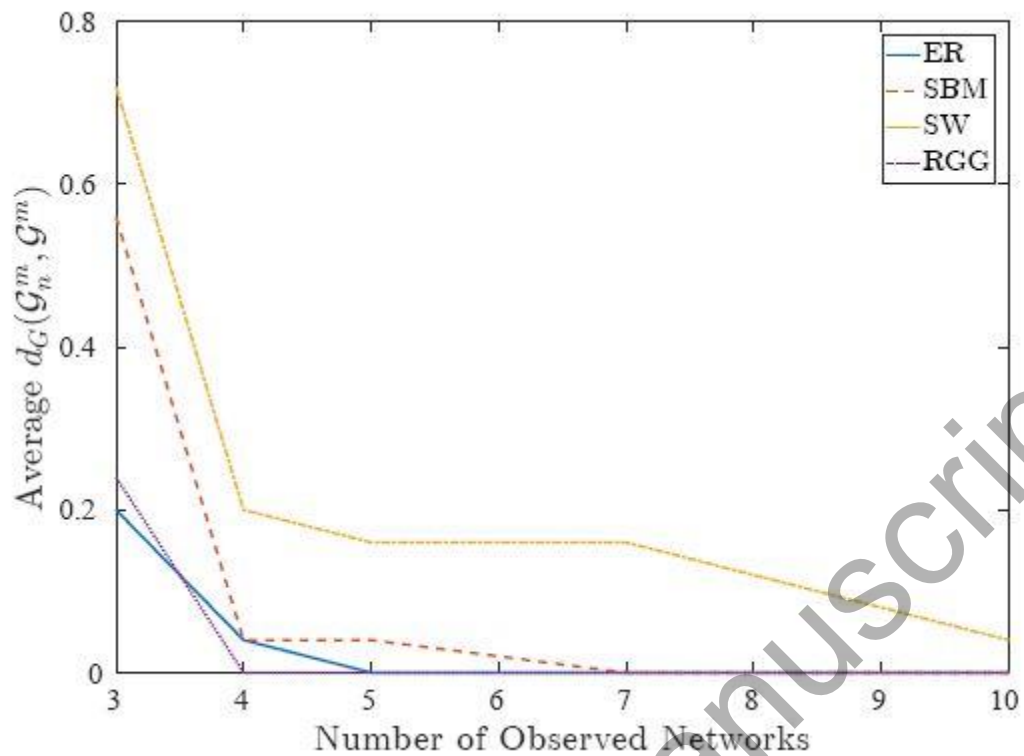
Fig. 3 Average distance of posterior mode to $\mathcal{G}^m$ as a function of sample size.
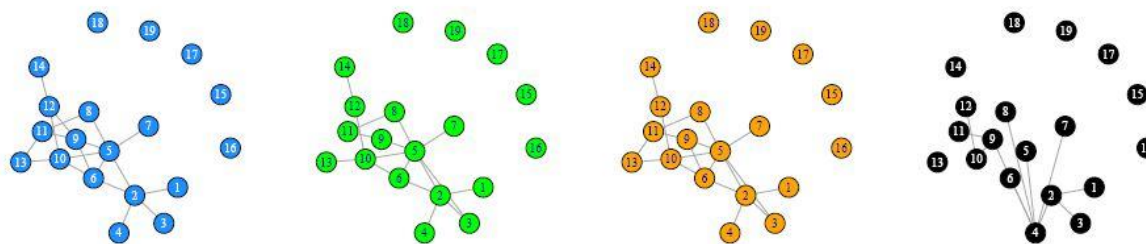
Fig. 4 Example of multiple network data in the context of cancer genomics, with each node one of the 19 most frequently mutated human cancer genes (see Section 6). Top Left: Network N_1 inferred from curated databases; Top Right: Network N_2 determined by a series of individual experiments; Bottom Left: Network N_3 inferred via text mining; Bottom Right: Network N_4 inferred via co-expression. The set of nodes of this network is formed by the 19 most frequently mutated human cancer genes.

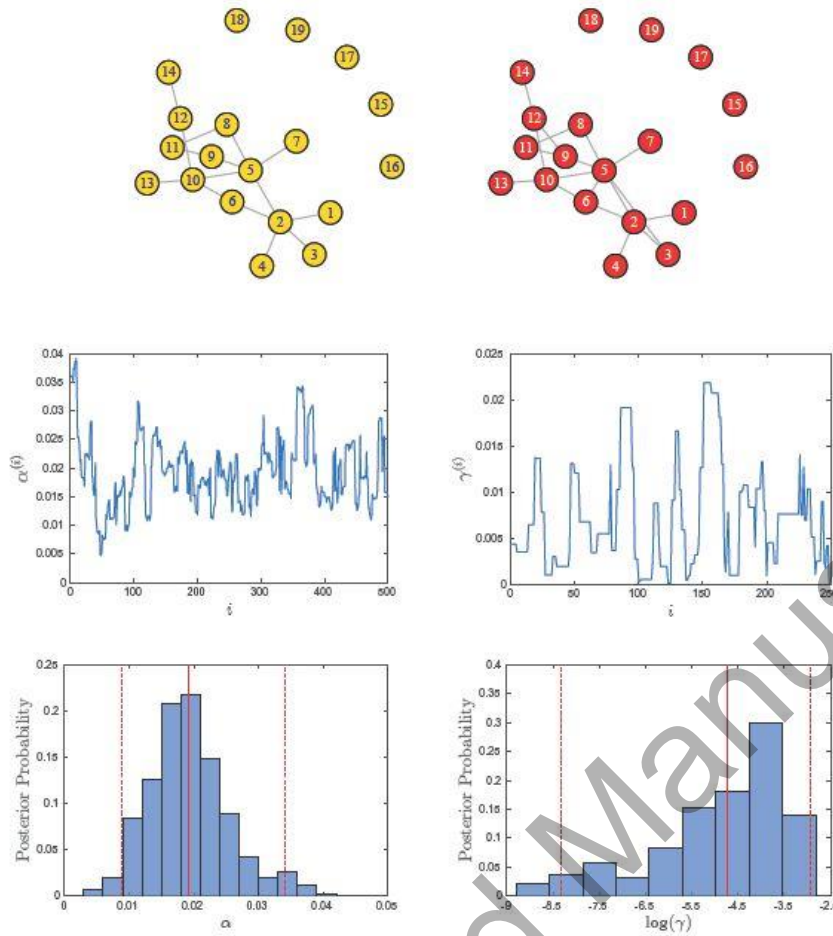**Fig. 5** Upper Left: Posterior mode from the SER/ SER model applied to the data set $\{N1, N2, N3\}$ and prior for the centroid centered at $N4$. Centre Left: Traceplot for 500 posterior samples for α after a burn-in of 150,000 and a lag of 50. Lower Left: Histogram for α. The posterior mean (highlighted by the red solid line) is equal to 0.0192. The 95% credible interval for α (delimeted by the dotted lines) is (0.0089,0.0342). Upper Right: Posterior mode obtained from fitting the SN/ SN model to the data set $\{N1, N2, N3\}$. This graph concentrates 0.544 of the posterior mass. Centre Right: Traceplot for 250 posterior samples for γ after a burn-in of 100,000 and a lag of 50. Lower Right: Histogram for $\log(\gamma)$. The posterior mean is equal to -4.6177. The 95% credible interval for $\log(\gamma)$ is (-8.4130,-2.9866).
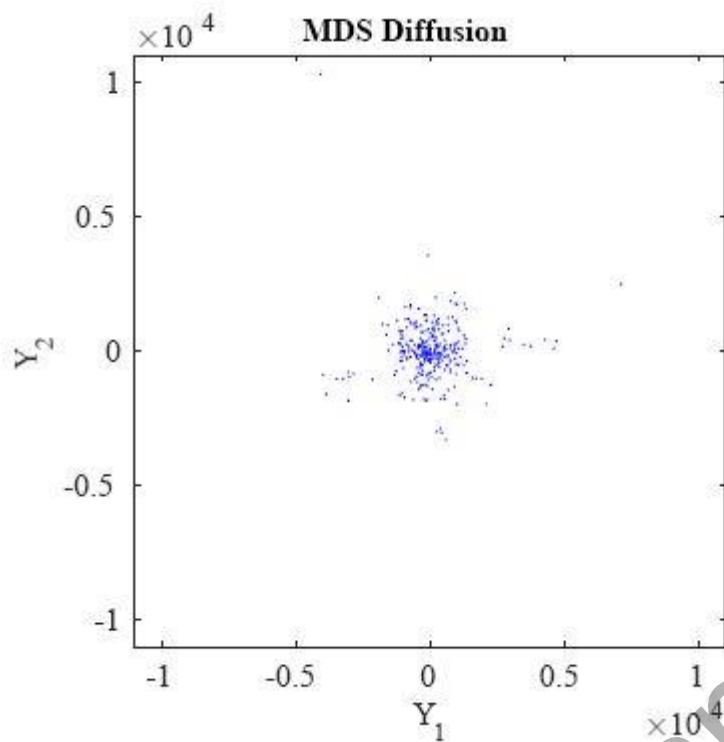
Fig. 6 The two-dimensional map obtained from applying multidimensional scaling on the 300 connectomes. The similarity is given by the diffusion distance.
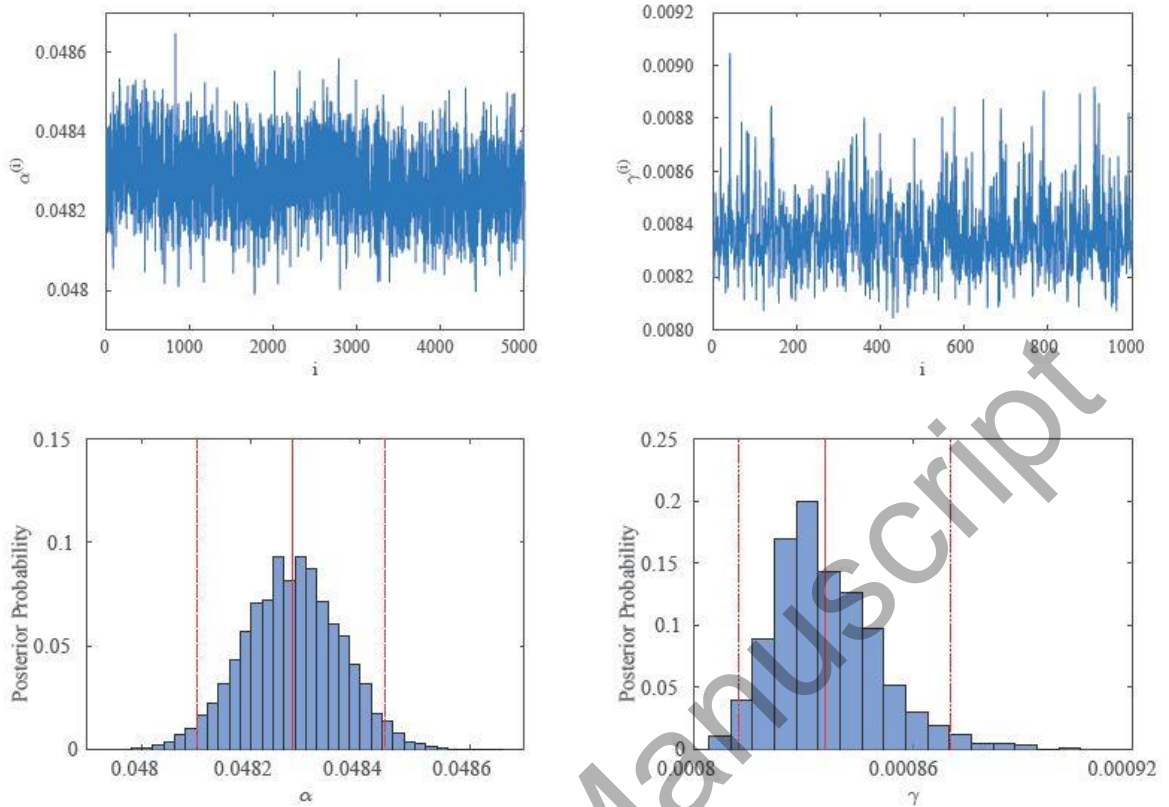
Fig. 7 Upper Left: Posterior mode from the SER/ SER model applied to the whole connectome data set and prior for $\mathcal{G}^m$ centered at the centroid of one of the subsets of the data. The posterior mode is presented in terms of its discrepancies with respect to the point estimate from the SN/ SN model. Edges only present in the posterior mode from SER/ SER are colored in blue, while edges present in the point estimate from the SN/ SN model but not the posterior mode are colored in pink. Centre Left: Traceplot for 5000 posterior samples for α after a burn-in of 150,000 and a lag of 50. Lower Left: Histogram for α. The posterior mean (highlighted by the red solid line) is equal to 0.0483. The 95% credible interval for α (delimeted by the dotted lines) is (0.0481,0.0485). Upper Right: Point estimate obtained from fitting the SN/ SN model to the full data set. Centre Right: Traceplot for 250 posterior samples for γ after a burn-in of 100,000 and a lag of 50. Lower Right: Histogram for $\log(\gamma)$. The posterior mean is equal to -4.6177. The 95% credible interval for $\log(\gamma)$ is (-8.4130,-2.9866).
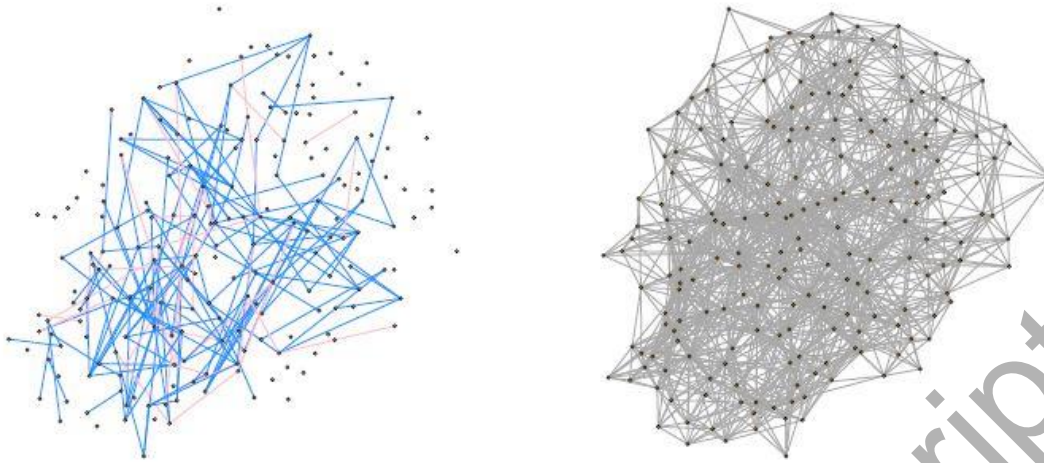
Fig. 8 Right: Point estimate for $\mathcal{G}^m$ obtained from computing the centroid of the posterior modes associated to each of the 10 subsets of the data. Each of these posterior modes was obtained from fitting the SN/ SN model. Left: Discrepancies of the posterior mode corresponding to one of the subsets of the data. Edges only present in the centroid are colored in blue, while edges present in the mode but not the centroid are colored in pink.

Table 1 Random graph models and the corresponding parameter specification used to define the simulation regimes. The parameters were chose so realisations would have approximately the same density across different models.

Random Graph Model	Specification
ER	Probability of inclusion was set to 0.1.
RGG	This is a proximity graph defined on the unit square.
	The radius of the ball was set to $r = 0.175$.
SBM	We set the number of blocks $K = 3$, with all membership
	probabilities equal to 0.333 the inclusion probabilities
	were set as 0.16 and 0.075 for diagonal and
	non-diagonal blocks, respectively.
SW	We set the degree of the lattice to 2 and the
	probability of re-wiring to 0.2.

Table 2 Proportion of replications where $1 - \delta$ of the posterior mass for $\mathcal{G}^m$ is within a ball of radius ϵ of the true value. We used 100 replications.

n	Generative Model for $\mathcal{G}^m$	$\epsilon = 1$	$\epsilon = 2$	$\epsilon = 3$
			$\delta = 0.05$	
3	RGG	0.92	1	1
5	RGG	1	1	1
3	ER	0.66	0.97	1
5	ER	0.93	1	1
7	ER	1	1	1
3	SBM	0.63	0.87	0.96
5	SBM	0.83	0.98	1
7	SBM	0.91	1	1
10	SBM	1	1	1
3	SW	0.43	0.61	0.73
5	SW	0.62	0.77	0.89
7	SW	0.74	0.86	0.98
10	SW	0.81	0.96	1

Table 3 Proportion of replications where $1-\delta$ of the posterior mass for $\mathcal{G}^m$ is within a ball of radius ϵ of the true value. We used 100 replications.

n	Generative Model for $\mathcal{G}^m$	$\epsilon = 1$	$\epsilon = 2$	$\epsilon = 3$	$\epsilon = 1$	$\epsilon = 2$	$\epsilon = 3$
			Majority Vote			CER	
3	RGG	0.91	1	1	0.94	0.99	1
5	RGG	1	1	1	1	1	1
3	ER	0.99	1	1	0.96	0.99	1
5	ER	1	1	1	1	1	1
3	SBM	0.95	0.99	1	0.95	0.99	1
5	SBM	1	1	1	1	1	1
3	SW	0.97	1	1	0.91	0.99	1
5	SW	1	1	1	0.99	1	1

Table 4 Average value ψ_δ for size of neighborhood needed so m samples from the predictive implied by n data points encloses $1-\delta$ of the predictive distribution associated with the true value of $\mathcal{G}^m$ and α. Here we assume a CER/ CER model (third column, left to right) and a SN/ SN model (fourth column, left to right). For the CER/ CER model $\rho_\delta = 17$, while for the SN/ SN model, $\rho_\delta = 3642.1$. The size of the network is 50 and $\alpha = 0.01$. We set $\delta = 0.1$ and $m = 20$ for all regimes. The average is computed over 100 replications.

n	Generative Model for $\mathcal{G}^m$	CER/ CER $\psi_\delta / \rho_\delta$	SN/ SN $\psi_\delta / \rho_\delta$
3	ER	1.4447	1.0551
5	ER	1.3847	1.0253
7	ER	1.3676	1.0072
10	ER	1.3612	0.9590
3	RGG	1.4006	1.0516
5	RGG	1.3988	1.0247
7	RGG	1.3953	0.9958
10	RGG	1.3635	0.9366
3	SBM	1.4141	1.0573
5	SBM	1.3824	1.0410
7	SBM	1.3800	0.9898
10	SBM	1.3741	0.9516
3	SW	1.4788	1.0697
5	SW	1.4494	1.0419
7	SW	1.3953	0.9937
10	SW	1.3682	0.9545

Table 5 Proportion of times where each diagnostic provided evidence for lack of fit over 100 simulated data sets. The regimes are given by the generative model, the type of misspecification and the univariate summary used for the diagnostics.

n	Model used to fit the data	type of misspecification	univariate summary	PPC	Bayes χ^2
3	Spherical Network Model	Dependence	10 quantile of degree	0	0
3	Spherical Network Model	Dependence	50 quantile degree	0	0
3	Spherical Network Model	Dependence	90 quantile degree	0.08	0
10	Spherical Network Model	Dependence	10 quantile of degree	0.02	0.02
10	Spherical Network Model	Dependence	50 quantile degree	0	0.01
10	Spherical Network Model	Dependence	90 quantile degree	0.09	0.11
50	Spherical Network Model	Dependence	10 quantile of degree	0.07	0.05
50	Spherical Network Model	Dependence	50 quantile degree	0.02	0.03
50	Spherical Network Model	Dependence	90 quantile degree	0.76	0.93
3	Spherical Network Model	Metric	10 quantile of degree	0.09	0
3	Spherical Network	Metric	50 quantile	0	0

n	Model used to fit the data	type of misspecification	univariate summary	PPC	Bayes χ^2
	Model		degree		
3	Spherical Network Model	Metric	90 quantile degree	0.11	0
10	Spherical Network Model	Metric	10 quantile of degree	0.03	0.07
10	Spherical Network Model	Metric	50 quantile degree	0.02	0.03
10	Spherical Network Model	Metric	90 quantile degree	0.05	0.07
50	Spherical Network Model	Metric	10 quantile of degree	0.07	0.22
50	Spherical Network Model	Metric	50 quantile degree	0	0.14
50	Spherical Network Model	Metric	90 quantile degree	0.09	0.97
3	Centred Erdös–Rényi Model	Dependence	10 quantile of degree	0	0
3	Centred Erdös–Rényi Model	Dependence	50 quantile degree	0	0
3	Centred Erdös–Rényi Model	Dependence	90 quantile degree	0.03	0.04
10	Centred Erdös–Rényi Model	Dependence	10 quantile of degree	0.02	0.01
10	Centred Erdös–Rényi Model	Dependence	50 quantile degree	0	0.00
1	Centred Erdös–Rényi	Dependence	90 quantile	0.0	0.14

n	Model used to fit the data	type of misspecification	univariate summary	PPC	Bayes χ^2
0	Model		degree	7	
50	Centred Erdös–Rényi Model	Dependence	10 quantile of degree	0.07	0.05
50	Centred Erdös–Rényi Model	Dependence	50 quantile degree	0.03	0.02
50	Centred Erdös–Rényi Model	Dependence	90 quantile degree	0.74	0.89
3	Centred Erdös–Rényi Model	Metric	10 quantile of degree	0.03	0
3	Centred Erdös–Rényi Model	Metric	50 quantile degree	0	0
3	Centred Erdös–Rényi Model	Metric	90 quantile degree	0.12	0
10	Centred Erdös–Rényi Model	Metric	10 quantile of degree	0.06	0.05
10	Centred Erdös–Rényi Model	Metric	50 quantile degree	0.03	0.01
10	Centred Erdös–Rényi Model	Metric	90 quantile degree	0.07	0.11
50	Centred Erdös–Rényi Model	Metric	10 quantile of degree	0.07	0.16
50	Centred Erdös–Rényi Model	Metric	50 quantile degree	0	0.12
50	Centred Erdös–Rényi Model	Metric	90 quantile degree	0.11	0.93

Table 6 Key for indices assigned to the 19 genes related to cancer.

Index	Gene	Index	Gene	Index	Gene
1	BRAF	8	PTEN	15	CIC
2	NRAS	9	CDKN2A	16	DNMT3A
3	ERBB3	10	CTNNB1	17	BFXW7
4	NF1	11	TP53	18	SF3B1
5	PIK3CA	12	SMAD4	19	LPHN2
6	PIK3R1	13	APC		
7	FLT3	14	NCOR1		

Table 7 The four networks with highest posterior mass obtained by fitting the CER/ CER model to the data set $\{N1, N2, N3\}$. Here $\mathcal{E}_{mode}$ denotes the edge set for the posterior mode.

Posterior Probability	Edge Set
0.246	1-2, 2-3, 2-4, 2-5, 2-6, 5-6, 5-7,
	5-8, 5-9, 5-10, 6-10, 8-11, 9-11,
	10-12, 10-13, 12-14
0.168	$\mathcal{E}_{mode} + (3\text{-}5)$
0.140	$\mathcal{E}_{mode} + (11\text{-}13)$
0.114	$\mathcal{E}_{mode} + (6\text{-}9)$

Table 8 The three networks with highest posterior mass obtained by fitting the SN/ SN model to the data set $\{N1, N2, N3\}$. Here $\mathcal{E}_{mode}$ denotes the edge set for the posterior mode.

Posterior Probability	Edge Set
0.544	1-2, 2-3, 2-4, 2-5, 2-6, 3-5, 5-6, 5-7,
	5-8, 5-9, 5-10, 6-10, 8-11, 9-11, 9-12,
	10-12, 10-13, 12-14
0.216	$\mathcal{E}_{mode} + (3,12) + (4,6) + (6,9) - (9,12)$
	$-(5,10)$
0.188	$\mathcal{E}_{mode} + (3,12) + (4,6) - (5,10)$