Modeling Network Populations via Graph Distances

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

View supplementary material

Published online: 08 Sep 2020.

Submit your article to this journal

View related articles

View Crossmark data

Taylor & Francis
Taylor & Francis Group

Modeling Network Populations via Graph Distances

Simón Lunagómez[a], Sofia C. Olhede[b,c], and Patrick J. Wolfe[d,e,f]

[a]Department of Mathematics and Statistics, Lancaster University, Lancaster, UK; [b]Institute of Mathematics, École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland; [c]Department of Statistical Science, UCL, London, UK; [d]Department of Statistics, Purdue University, West Lafayette, IN; [e]Department of Computer Science, Purdue University, West Lafayette, IN; [f]Department of Electrical & Computer Engineering, Purdue University, West Lafayette, IN

ABSTRACT

This article introduces a new class of models for multiple networks. The core idea is to parameterize a distribution on labeled 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. This article has online supplementary materials.

1. 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 datasets or to different analysis procedures applied to the same data (Bartlett, Olhede, and Zaikin 2014). In neuroscience, a population of networks encodes the way different regions of the brain interact when individuals perform a given task (Biswal, Menness, and Zuo 2010), or characterizes a population of individuals suffering from a neurological or psychiatric disorder (Lynall, Bassett, and Kerwin 2010; Nelson et al. 2017).

The developments proposed herein are therefore motivated by the problem of modeling populations of networks. The class of models we propose is based on the idea that distributions on graph space are naturally parameterized 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 modeler 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 nontrivial 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 Figure 1, with regions of the brain assigned to nodes according to the CC200 atlas, which was proposed by Craddock et al. (2012). Note from Figure 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 (Figure 1, on the right). Thus, although one modeling approach would be to treat such networks as realizations 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 modeler impose on the space of labeled graphs to define a suitable estimand? Without such an estimand (e.g., in the case of block modeling 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.
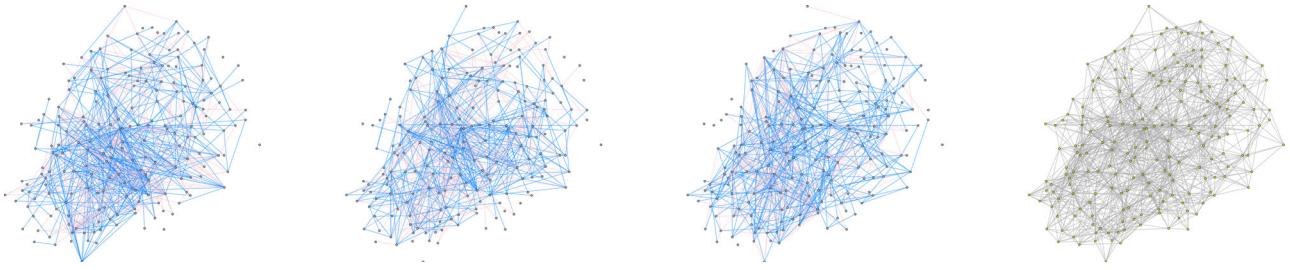
Figure 1. Example of multiple network data in the context of neuroscience, with each node representing a region of the brain (see Zuo et al. 2014; Arroyo et al. 2019). The networks are defined over the same set of 200 nodes. First three figures (from left to right): Discrepancies of three observed brain 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.

2. In the Bayesian setting, if we have multiple networks (such as those in Figure 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 characterizing the variability of the data in those terms. Specifically, the Fréchet mean implied by a given metric will parameterize 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, Kolaczyk, and Viles (2017) and Chang, Kolaczyk, and Yao (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, Raftery, and Handcock (2002)) for describing the variability of a homogeneous population of networks; the Bayesian nonparametric approach proposed by Durante, Dunson, and Vogelstein (2017) for modeling 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 unlabeled networks developed by Kolaczyk et al. (2020), 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. (2020) and earlier Feragen et al. (2011) discussed 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, Vogelstein, and Priebe (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) focused on the problem of comparing populations of networks. Tang et al. (2017) focused 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, Vogelstein, and Priebe (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 modeling measurement error for networks, including recent work by Newman (2018), Peixoto (2018b), and Le, Levin, and Levina (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 proposed 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 modeling 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 minimizing the posterior expected loss, and becomes the same problem when the loss is a metric. Wade and Ghahramani (2018) exploited 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 modeler to characterize 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 modeling 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, Kotz, and Ng (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 organized 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 article are be based on, along some examples. The corresponding strategies for Bayesian modeling and computation are presented in Section 4. Section 5 documents the behavior of our models via simulation studies, and Section 6 describes fully the fitting of our models to the multiple-network data introduced in Figure 1. Finally, Section 7 discusses briefly the contributions of our approach, placing it in context and outlining limitations as well as future possibilities.

2. Preliminaries

A simple labeled 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 $\mathcal{V} := \{1, 2, \ldots, N\}$, define

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

so that $\{\mathcal{G}_{[N]}\}$ represents the set of all N-node labeled networks of a given type (simple, directed, etc.). If we consider simple graphs, for example, then $|\{\mathcal{G}_{[N]}\}| = 2^{\binom{N}{2}}$. We refer to $\{\mathcal{G}_{[N]}\}$ 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 $\{\mathcal{G}_{[N]}\}$ given the choice of a metric $d_G(\cdot, \cdot)$ on $\{\mathcal{G}_{[N]}\}$. Two examples of metrics which can be used to formulate the models introduced in Section 3 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 modeler (e.g., simple, directed).

2. A *diffusion distance* based on the graph Laplacian, for example the choice made by Hammond, Gur, and Johnson (2013):

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

where $\|\cdot\|_F$ 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 Chung (1997), 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, Gur, and Johnson (2013) 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 modeled 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, Gur, and Johnson (2013) argued 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) provided 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) discussed, 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). Computational 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 $\left(\left\{\mathcal{G}_{[N]}\right\}, d_G\right)$ 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 $\left\{\mathcal{G}_{[N]}\right\}$ of labeled N-node networks.

3. Modeling Approach

In this section, we propose a generative modeling approach for datasets consisting of multiple networks. Our models are parameterized in terms of a unique mode and a univariate measure of dispersion around that mode. The mode in the space of labeled N-node networks $\left\{\mathcal{G}_{[N]}\right\}$ 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 $\left\{\mathcal{G}_{[N]}\right\}$ 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 centered Erdös–Rényi (CER) 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

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

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 and scale). Fix a metric $d_G(\cdot, \cdot)$ on $\left\{\mathcal{G}_{[N]}\right\}$ for $N \in \mathbb{N}$, nonempty set $\Gamma \subset \mathbb{R}_+$, and consider a family $\left\{p(\cdot \mid \mathcal{G}^m, \gamma)\right\}_{\mathcal{G}^m \in \left\{\mathcal{G}_{[N]}\right\}, \gamma \in \Gamma}$:

1. For every fixed scale parameter $\gamma^* \in \Gamma$, the family $\left\{p(\cdot \mid \mathcal{G}^m, \gamma^*)\right\}_{\mathcal{G}^m \in \left\{\mathcal{G}_{[N]}\right\}}$ satisfies Definition 3.1 with respect to the metric d_G.
2. For every fixed location parameter $\mathcal{G}^* \in \left\{\mathcal{G}_{[N]}\right\}$, the entropy associated to the family $\left\{p(\cdot \mid \mathcal{G}^*, \gamma)\right\}_{\gamma \in \Gamma}$ is a strictly monotone function of $\gamma \in \Gamma$.

For finite, discrete sets such as $\left\{\mathcal{G}_{[N]}\right\}$ and associated probability mass function $p(\cdot)$, entropy $-\mathbb{E}\left\{\log p(\cdot)\right\}$ provides a convenient characterization akin to variance, ranging from 0 for a point mass to $\log(|\left\{\mathcal{G}_{[N]}\right\}|)$ for the uniform distribution on $\left\{\mathcal{G}_{[N]}\right\}$. Entropy can thus be used to parameterize a family of discrete distributions on $\left\{\mathcal{G}_{[N]}\right\}$ with the same unique mode, in an analogous way to how the scale parameter would parameterize a member of the location–scale family when the location parameter has been specified. The metric provides a ranking of the elements $\left\{\mathcal{G}_{[N]}\right\}$ 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 CER 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 modeling 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 (CER model). Given a graph $\mathcal{G}^m \in \left\{\mathcal{G}_{[N]}\right\}$ and $1/2 > \alpha > 0$, consider a model $p(\cdot \mid \mathcal{G}^m, \alpha)$ 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 CER model 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 Ber(α) for $1 \leq i < j \leq N$.

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 $\alpha \leq \frac{1}{2}$, 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 (online appendix), which exactly establishes mode etc. Thus, for the CER to effectively serve as an error measurement model, the α parameter should be constrained to 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 minimizes 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 Newman (2018).

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 CER model with centroid $\mathcal{G}^m \in \{\mathcal{G}_{[N]}\}$ and $0 < \alpha \leq 1/2$ 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 CER 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 Zelinka (1975), Hammond, Gur, and Johnson (2013), or the ones discussed in Donnat and Holmes (2018)), and covered by our discussion in Section 2 earlier in the article.

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 nonnegative strictly increasing function such that $\phi(0) = 0$. This is the spherical network family (SNF) with parameters $\mathcal{G}^m$ and γ.

This model is related to the prior introduced by Mitra et al. (2013), which was introduced in the context of graphical modeling. A main difference with respect to their approach is that the SNF 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 Dahl, Day, and Tsai (2017) 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 SNF is inspired by the notion of symmetry of the density discussed in Fang, Kotz, and Ng (1990). 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 nondecreasing 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 SNF 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 SNF is a unimodal network distribution based on location. In addition, the SNF is unimodal network distribution based on location and scale if $\mathbb{V}\text{ar}\{\phi[d(\mathcal{G}, \mathcal{G}^m)]\} > 0$.

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 $\mathbb{V}\text{ar}\{\phi[d(\mathcal{G}, \mathcal{G}^m)]\} > 0$ 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 CER model:

Proposition 3.6. The mode and Fréchet mean coincide for the CER model defined on $\{\mathcal{G}_{[N]}\}$, $N \in \mathbb{N}$.

For the SNF, the Fréchet mean maximizes the kernel of the Boltzmann distribution in Equation (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 $\{\mathcal{G}_{[N]}\}$. To describe the variability of a set observations $\{\mathcal{G}_1, \mathcal{G}_2, \ldots, \mathcal{G}_n\}$ in $\{\mathcal{G}_{[N]}\}^n$, we propose a model of the form:

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

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 parameterization 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 $\{\mathcal{G}_1, \mathcal{G}_2, \ldots, \mathcal{G}_n\}$ in $\{\mathcal{G}_{[N]}\}^n$ such that, for every pair $\{\mathcal{G}_i, \mathcal{G}_j\}$ with $1 \le i < j \le n$, $\mathcal{G}_i$ 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.
2. The variability of the observations is characterized 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.

4. Bayesian Modeling 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 normalizing constant depends on the Fréchet mean.

4.1. Bayesian Inference for the Centered Erdös–Rényi Model

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

$$p(\mathcal{G}_1, \mathcal{G}_2, \ldots, \mathcal{G}_n \mid \mathcal{G}^m, \alpha) \quad (8)$$
$$= \alpha_0^{d_H(\mathcal{G}^m, \mathcal{G}_0)} (1 - \alpha_0)^{\frac{(N-1)N}{2} - d_H(\mathcal{G}^m, \mathcal{G}_0)} p(\alpha)$$
$$\times \prod_{i=1}^{n} \alpha^{d_H(\mathcal{G}_i, \mathcal{G}^m)} (1 - \alpha)^{\frac{(N-1)N}{2} - d_H(\mathcal{G}_i, \mathcal{G}^m)},$$

where the prior $p(\cdot)$ for α is a scaled Beta on $(0, \frac{1}{2})$. Here, $\mathcal{G}_0 \in \{\mathcal{G}_{[N]}\}$ 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 favors 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 $\widehat{\mathcal{G}^m}$ that minimizes 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 Ber $\left\{\frac{1}{n}\sum_{k=1}^{n}A_{\mathcal{G}_k}(i,j)\right\}$.

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 $\zeta \sim \text{Unif}(-\upsilon_k, \upsilon_k)$, if $0 < y < 0.5$;
2. $-y$, if $y < 0$;
3. $1 - y$, if $y > 0.5$.

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

4.2. Bayesian Inference for Models in the SNF

The SNF 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 $\{\mathcal{G}_{[N]}\}$. The SN/SN model is a multivariate network model on $\{\mathcal{G}_{[N]}\}^n$ of the form

$$p(\mathcal{G}_1, \mathcal{G}_2, \ldots, \mathcal{G}_n \mid \mathcal{G}^m, \gamma) \qquad (9)$$
$$\propto \exp\left\{-\gamma_0 \phi(d_G(\mathcal{G}^m, \mathcal{G}_0))\right\} p(\gamma)$$
$$\times \exp\left\{-\gamma \sum_{i=1}^{n} \phi(d_G(\mathcal{G}_i, \mathcal{G}^m))\right\},$$

where $p(\cdot)$ is the prior on γ, which has support on $\mathbb{R}^+$. Here, $\mathcal{G}_0 \in \{\mathcal{G}_{[N]}\}$ 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, for example, 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:

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

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

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

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

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 $\mathcal{G}_i$, $1 \leq i \leq n$. 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:

$$H_{(\mathcal{G}^{m,(t+1)}, \gamma^{(t+1)} \mid \mathcal{G}^{m,(t)}, \gamma^{(t)})} \qquad (12)$$
$$= \frac{f(\vec{\mathcal{G}}_*^{(t+1)} \mid \mathcal{G}^{m,(t+1)}, \tilde{\alpha})}{f(\vec{\mathcal{G}}_*^{(t)} \mid \mathcal{G}^{m,(t)}, \tilde{\alpha})} \times \frac{p(\mathcal{G}^{m,(t+1)} \mid \mathcal{G}_0, \gamma_0)}{p(\mathcal{G}^{m,(t)} \mid \mathcal{G}_0, \gamma_0)}$$
$$\times \frac{p(\vec{\mathcal{G}} \mid \mathcal{G}^{m,(t+1)}, \gamma^{(t+1)})}{p(\vec{\mathcal{G}} \mid \mathcal{G}^{m,(t)}, \gamma^{(t)})} \times \frac{p(\vec{\mathcal{G}}_*^{(t)} \mid \mathcal{G}^{m,(t)}, \gamma^{(t)})}{p(\vec{\mathcal{G}}_*^{(t+1)} \mid \mathcal{G}^{m,(t+1)}, \gamma^{(t+1)})}$$
$$\times \frac{q(\mathcal{G}^{m,(t)}, \gamma^{(t)} \mid \mathcal{G}^{m,(t+1)}, \gamma^{(t+1)})}{q(\mathcal{G}^{m,(t+1)}, \gamma^{(t+1)} \mid \mathcal{G}^{m,(t)}, \gamma^{(t)})},$$

where the terms of the form:

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

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

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

2. $p(\vec{\mathcal{G}}_*^{(\cdot)} \mid \mathcal{G}^{m,(\cdot)}, \gamma^{(\cdot)})$ 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. $p(\mathcal{G}^{m,(\cdot)}, \gamma^{(\cdot)} \mid \mathcal{G}_0, \gamma_0)$ correspond to the prior for $(\mathcal{G}^m, \gamma)$ evaluated at the state of the chain.

4. $q(\mathcal{G}^{m,(\cdot)}, \gamma^{(\cdot)} \mid \mathcal{G}^{m,(\cdot)}, \gamma^{(\cdot)})$ 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. $f(\vec{\mathcal{G}}_*^{(\cdot)} \mid \mathcal{G}^{m,(\cdot)}, \tilde{\alpha})$ correspond to the conditional density of the auxiliary variables. We adopted the probability mass
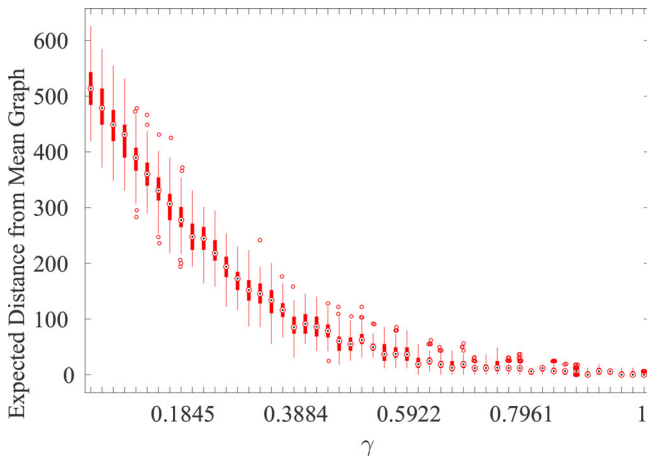
Figure 2. Distribution of the distance (summarized as a boxplot) to $\mathcal{G}^m$ for the SN model as a function of γ. Here, $N = 19$, $t = 1$, and $\mathcal{G}^m$ was specified as the network displayed in Figure 4. The figure shows that, for $\gamma > 1$ increments on that parameter do not have a perceptible effect on the distribution. This plot serves to inform the scales that are relevant for defining random walk proposals for γ.

function of the CER model as the conditional density for the auxiliary variables, which are denoted by $(\mathcal{G}_*^1, \ldots, \mathcal{G}_*^n)$, that is,

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

as in Møller et al. (2006, sec. 2). 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 Møller et al. (2006, eq. (7)).

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 $\zeta \sim \text{Unif}(-\upsilon_k, \upsilon_k)$, if $0 < y$;
2. $-y$, if $y < 0$;

The mixture is over $\{\upsilon_1, \upsilon_2, \ldots, \upsilon_K\}$. Here υ is a tuning parameter. Specifying a prior for γ also presents a challenge, since its behavior 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.

5. Simulation Studies

In this section, we explore the behavior 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 analyzed (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.

5.1. 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

$$\Pr\{d_G(\mathcal{G}, \mathcal{G}^m) > \epsilon \mid \mathcal{G}_1, \mathcal{G}_2, \ldots, \mathcal{G}_n\} \to 0, \qquad (13)$$

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 Ghosal and van der Vaart (2017, sec. 13.4.1). 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 neighborhood $\epsilon > 0$. 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 $\{\mathcal{G}_1, \mathcal{G}_2, \ldots, \mathcal{G}_n\}$. Equation (13) provides the principle behind the following simulation experiments:

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

$$\Pr\{d_G(\mathcal{G}, \mathcal{G}^m) > \epsilon \mid \mathcal{G}_1, \mathcal{G}_2, \ldots, \mathcal{G}_n\} < \delta, \qquad (14)$$

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 Figure 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
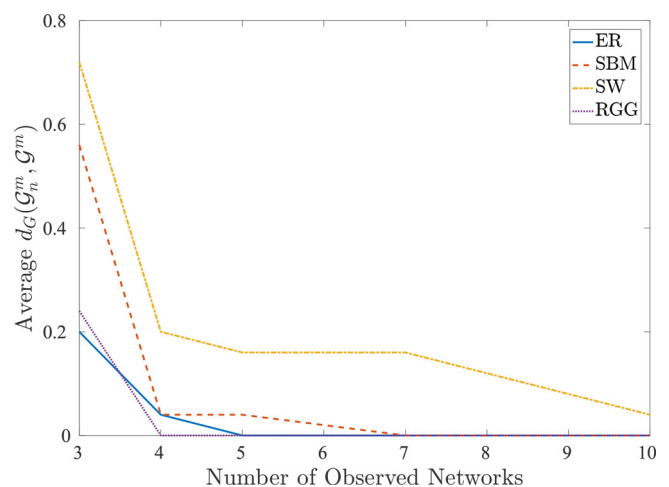

Figure 3. Average distance of posterior mode to $\mathcal{G}_m$ as a function of sample size for different random graph models as indicated by the legend.

Table 1. Random graph models and the corresponding parameter specification used to define the simulation regimes.

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 nondiagonal blocks, respectively.
SW	We set the degree of the lattice to 2 and the probability of rewiring to 0.2.

NOTE: The parameters were chose so realizations would have approximately the same density across different models.

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.

n	Generative model for $\mathcal{G}^m$	$\delta = 0.05$		
		$\epsilon = 1$	$\epsilon = 2$	$\epsilon = 3$
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

NOTE: We used 100 replications.

(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 summarized in Figure 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 rewiring on a lattice, it takes a larger sample size to disambiguate between the outcomes of the rewiring process and the perturbation induced by the SN model.

We compared the performance of our method to the point estimate $\widehat{\mathcal{G}^m}$ 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 summarized in Table 3.

5.2. Network Prediction

In this section, we investigate the behavior 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

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

for $(\mathcal{G}^m, \gamma)$ specified, and then we partition this sample into a training set $\{\mathcal{G}_i\}_{i \leq n}$ and a test set $\{\mathcal{G}_i\}_{n < i \leq n_t}$. 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

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.

n	Generative model for $\mathcal{G}^m$	Majority vote			CER		
		$\epsilon = 1$	$\epsilon = 2$	$\epsilon = 3$	$\epsilon = 1$	$\epsilon = 2$	$\epsilon = 3$
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

NOTE: We used 100 replications.

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

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

holds. In Equation (15), $\mathbb{B}(\mathcal{G}_k; \psi)$ denotes the ball with center $\mathcal{G}_k$ 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

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

holds, that is, 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 summarized 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 summarized 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.

5.3. 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, Meng, and Stern 1996), and (ii) by investigating the behavior of the Bayesian χ^2 (Johnson 2004) under different scenarios. These methods are further discussed in online 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 labeled 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, Gur, and Johnson 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 ($\mathcal{G}_0, \gamma_0$ and α_0) as in Sections 5.1 and 5.2. The size of the networks was set to $N = 50$.

Results are summarized in Table 5. In this table, we display the proportion of times where each diagnostic provided evidence for lack of fit over 100 simulated datasets. 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.

6. Data Analysis

6.1. 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 signaling) or other evidence of association. Estimating the network is often an intermediate step within a series of

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.

n	Generative model for $\mathcal{G}^m$	CER/CER ψ_δ/ρ_δ	SN/SN ψ_δ/ρ_δ
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

NOTE: 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.

Table 5. Proportion of times where each diagnostic provided evidence for lack of fit over 100 simulated datasets.

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 model	Metric	50 quantile degree	0	0
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	Centered Erdős–Rényi model	Dependence	10 quantile of degree	0	0
3	Centered Erdős–Rényi model	Dependence	50 quantile degree	0	0
3	Centered Erdős–Rényi model	Dependence	90 quantile degree	0.03	0.04
10	Centered Erdős–Rényi model	Dependence	10 quantile of degree	0.02	0.01
10	Centered Erdős–Rényi model	Dependence	50 quantile degree	0	0.00
10	Centered Erdős–Rényi model	Dependence	90 quantile degree	0.07	0.14
50	Centered Erdős–Rényi model	Dependence	10 quantile of degree	0.07	0.05
50	Centered Erdős–Rényi model	Dependence	50 quantile degree	0.03	0.02
50	Centered Erdős–Rényi model	Dependence	90 quantile degree	0.74	0.89
3	Centered Erdős–Rényi model	Metric	10 quantile of degree	0.03	0
3	Centered Erdős–Rényi model	Metric	50 quantile degree	0	0
3	Centered Erdős–Rényi model	Metric	90 quantile degree	0.12	0
10	Centered Erdős–Rényi model	Metric	10 quantile of degree	0.06	0.05
10	Centered Erdős–Rényi model	Metric	50 quantile degree	0.03	0.01
10	Centered Erdős–Rényi model	Metric	90 quantile degree	0.07	0.11
50	Centered Erdős–Rényi model	Metric	10 quantile of degree	0.07	0.16
50	Centered Erdős–Rényi model	Metric	50 quantile degree	0	0.12
50	Centered Erdős–Rényi model	Metric	90 quantile degree	0.11	0.93

NOTE: The regimes are given by the generative model, the type of misspecification and the univariate summary used for the diagnostics.

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 characterization 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 preprocess 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 fulfills 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 Figure 5. The edge sets corresponding to four networks with highest posterior probability are displayed in Table 7. The posterior mode is displayed in Figure 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 dataset 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 Figure 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, Gur, and Johnson 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 Figure 5(upper
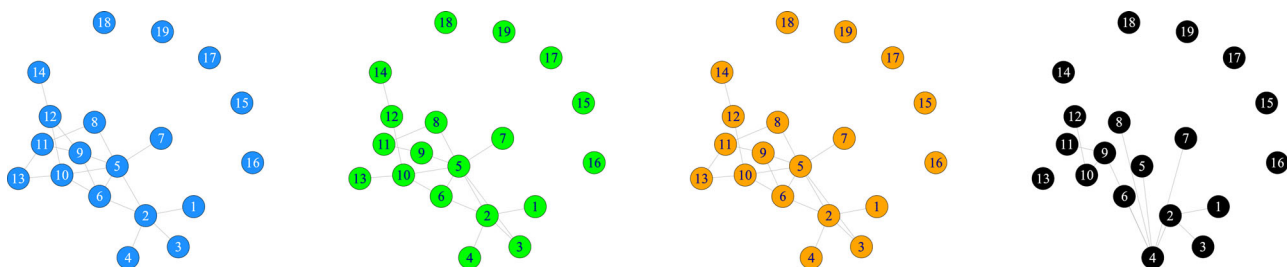
Figure 4. Example of multiple network data in the context of cancer genomics, with each node one of the 19 most frequently mutated human cancer genes (see Section 6). Far left: Network N1 inferred from curated databases; left middle: network N2 determined by a series of individual experiments; right middle: network N3 inferred via text mining; far right: network N4] inferred via co-expression. The set of nodes of this network is formed by the 19 most frequently mutated human cancer genes .

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 dataset {N1, N2, N3}.

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	$\mathcal{E}_{mode}$ + (3–5)
0.140	$\mathcal{E}_{mode}$ + (11–13)
0.114	$\mathcal{E}_{mode}$ + (6–9)

NOTE: Here $\mathcal{E}_{mode}$ denotes the edge set for the posterior mode.

Table 8. The three networks with highest posterior mass obtained by fitting the SN/SN model to the dataset {N1, N2, N3}.

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)

NOTE: Here, $\mathcal{E}_{mode}$ denotes the edge set for the posterior mode.

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 Figure 5. The choice of the metric penalizes 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 mathematical convenience (SER / SER model).

6.2. 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 consist of 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
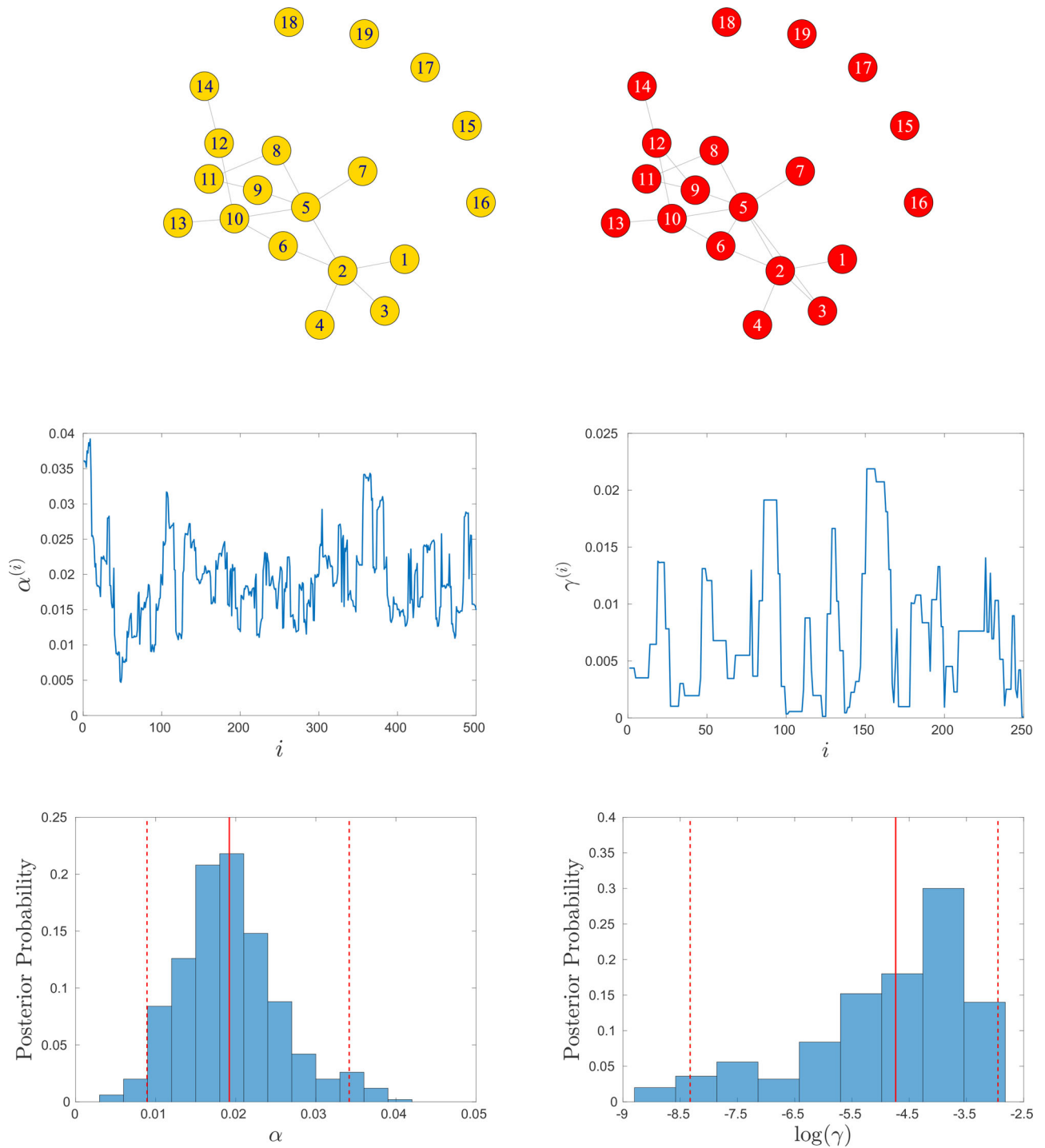
Figure 5. Top left: Posterior mode from the SER/SER model applied to the dataset {N1, N2, N3}; this prior concentrates 0.246 of the posterior mass. The prior for the centroid was centered at N4. Middle left: Traceplot for 500 posterior samples for α after a burn-in of 150,000 and a lag of 50. Bottom left: Histogram of the posterior samples for α. The posterior mean (highlighted by the red solid line) is equal to 0.0192. The 95% credible interval for α (delimited by the dotted lines) is (0.0089, 0.0342). Top right: Posterior mode obtained from fitting the SN/SN model to the dataset {N1, N2, N3}. This graph concentrates 0.544 of the posterior mass. Middle right: Traceplot for 250 posterior samples for γ after a burn-in of 100,000 and a lag of 50. Bottom right: Histogram for $\log(\gamma)$. The posterior mean is equal to -4.6177. The 95% credible interval for $\log(\gamma)$ is $(-8.4130, -2.9866)$.

of labeled 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 modeling 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 dataset, we used 2,000,000 iterations for burn-in and obtained 5000 samples with a lag of 1000. In contrast, fitting the SN/SN model for a dataset is not straightforward. We followed the divide-and-conquer
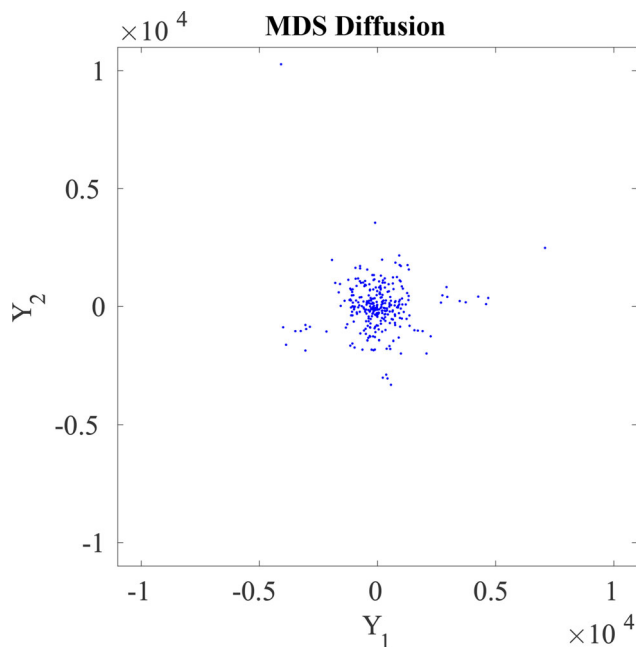
Figure 6. The two-dimensional map obtained from applying multidimensional scaling on the 300 connectomes. The similarity is given by the diffusion distance.

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 parameterized 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 dataset. 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 parameterized so γ is on the same scale across the ten subsets. We only need to: (i) recenter the all samples with respect to the sample mean of the corresponding subset; (ii) rescale so each posterior has roughly the same dispersion as the full posterior; (iii) recenter again using the global sample mean.

For each subset, we ran the MCMC described in Section 4.2. We used 500,000 iterations for burn-in and obtained 1000 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 datasets 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 dataset (Figure 8).

7. 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 article have designed a modular framework that allows us to specify each component, and thus to model.

This modular framework can be compared to the modeling framework of others, such as Newman (2018), Le, Levin, and Levina (2018), Chang, Kolaczyk, and Yao (2018), Durante, Dunson, and Vogelstein (2017), and Peixoto (2018a). In comparison to Durante, Dunson, and Vogelstein (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, Levin, and Levina (2018), Chang, Kolaczyk, and Yao (2018), and Peixoto (2018a), by contrast, our parameterizations are more complex while we adopt a similarly simple characterization of perturbations from the typical network.

The use of the Fréchet mean as a parameter that encodes what the center 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 modeling the variability of a homogeneous population of shapes are listed as two of the main challenges in that area in Srivastava and Klassen (2016, sec. 1.3). We pose these challenges in the context of network data and offer solutions for the implied inference problems via Bayesian modeling. 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 prespecified 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
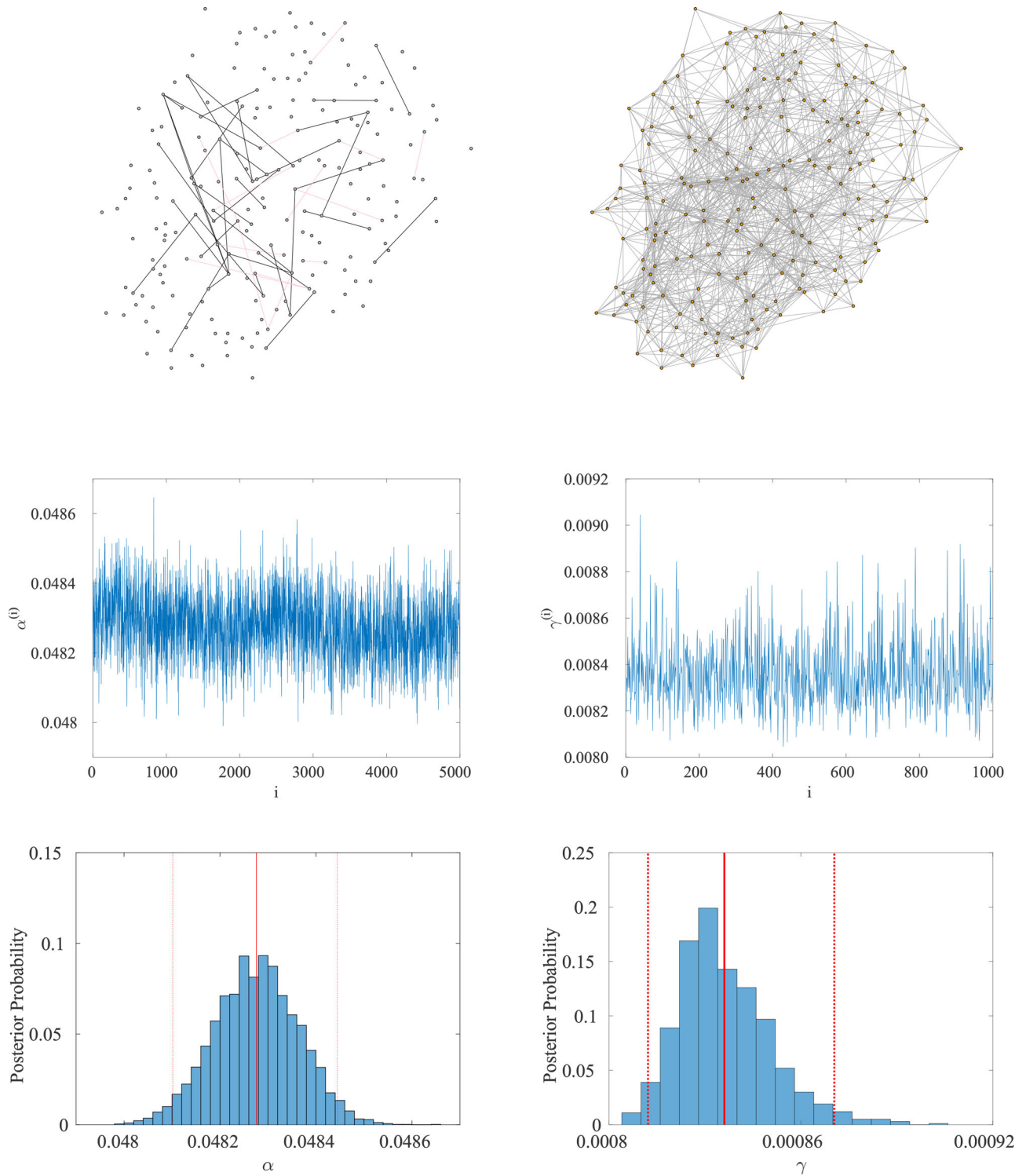
Figure 7. Upper left: Posterior mode from the SER/SER model applied to the whole connectome dataset and prior for $\mathcal{G}^m$ centered at the centroid of one of the subsets of the data. The posterior mode is presented in terms of its discrepancies with respect to the point estimate from the SN/SN 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. Center 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 (delimited by the dotted lines) is (0.0481, 0.0485). Upper right: Point estimate obtained from fitting the SN/SN model to the full dataset. Center 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 $(-8.4130, -2.9866)$.

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^2 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, Wu, and Xu 2017). Here, we see identifiability starts to depend on the number of nodes, and the observed
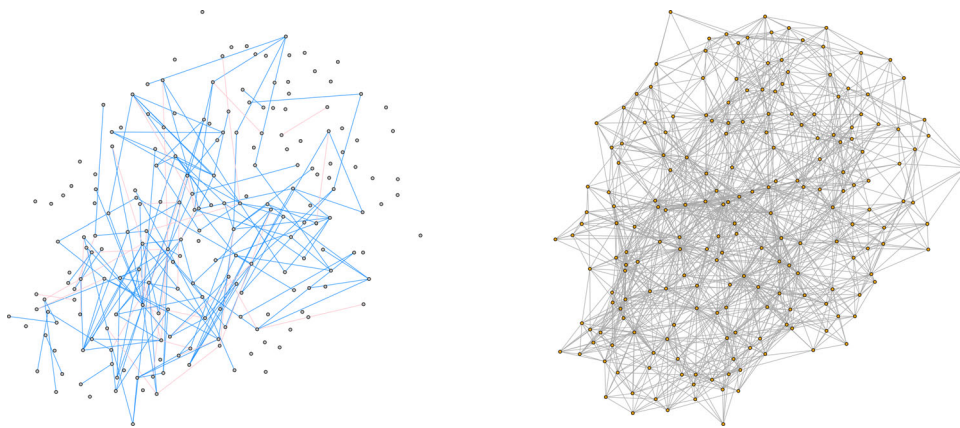
Figure 8. Right: Point estimate for $\mathcal{G}^m$ obtained from computing the centroid of the posterior modes associated to each of the 10 subsets of the data . Each of these posterior modes was obtained from fitting the SN/SN model. Left: Discrepancies of the posterior mode corresponding to one of 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.

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 article are still valid (provided an appropriate metric is provided) and the theory results will hold. In terms of computation, an MCMC based on a saturated model approach (as the one proposed in Brooks, Giudici, and Roberts (2003, sec. 5)) 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, Dunson, and Vogelstein (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 center, but derive the uncertainty around that center 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, Dunson, and Vogelstein (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 Newman 2018; Peixoto 2018b), 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.

Acknowledgments

Funding

References

Andrieu, C., and Roberts, G. O. (2009), "The Pseudo-Marginal Approach for Efficient Monte Carlo Computations," *The Annals of Statistics*, 37, 697–725. [7]

plsu.01pt minus.01pt

Arroyo, J., Athreya, A., Cape, J., Chen, G., Priebe, C. E., and Vogelstein, J. T. (2019), "Inference for Multiple Heterogeneous Networks With Common Invariant Subspace," arXiv no. 1906.10026. [2,12]

Balachandran, P., Kolaczyk, E. D., and Viles, W. D. (2017), "On the Propagation of Low-Rate Measurement Error to Subgraph Counts in Large Networks," *Journal of Machine Learning and Research*, 18, 1–33. [2]

Bartlett, T. E., Olhede, S. C., and Zaikin, A. (2014), "A DNA Methylation Network Interaction Measure and Detection of Network Oncomarkers," *PLoS One*, 9, e84573. [1]

Bhattacharyya, S., and Chatterjee, S. (2018), "Spectral Clustering for Multiple Sparse Networks: I," arXiv no. 1805.10594. [2]

Biswal, B. B., Menness, M., and Zuo, X.-N. (2010), "Toward Discovering Science of Human Brain Function," *Proceedings of the National Academy of Sciences of the United States of America*, 107, 4734–4739. [1]

Brooks, S. P., Giudici, P., and Roberts, G. O. (2003), "Efficient Construction of Reversible Jump Markov Chain Monte Carlo Proposal Distributions," *Journal of the Royal Statistical Society*, Series B, 65, 3–39. [16]

Chang, J., Kolaczyk, E. D., and Yao, Q. (2018), "Estimation of Edge Density in Noisy Networks," arXiv no. 1803.02488v1. [2,14]

Chung, F. K. R. (1997), *Spectral Graph Theory*, Providence, RI: American Mathematical Society. [3]

Craddock, R. C., James, G., Holtzheimer, P. E., Hu, X. P., and Mayberg, H. S. (2012), "A Whole Brain FMRI Atlas Generated via Spatially Constrained Spectral Clustering," *Human Brain Mapping*, 33, 1914–1928. [1,12]

Dahl, D. B., Day, R., and Tsai, J. W. (2017), "Random Partition Distribution Indexed by Pairwise Information," *Journal of the American Statistical Association*, 112, 721–732. [5]

Donnat, C., and Holmes, S. (2018), "Tracking Network Dynamics: A Survey Using Graph Distances," *The Annals of Applied Statistics*, 12, 971–1012. [3,4,5,12,13]

Dryden, I. L., and Mardia, K. V. (1998), *Statistical Shape Analysis*, Chichester: Wiley. [14]

Durante, D., Dunson, D. B., and Vogelstein, J. T. (2017), "Nonparametric Bayes Modeling of Populations of Networks," *Journal of the American Statistical Association*, 112, 1516–1530. [2,14,16]

Fang, K. T., Kotz, S., and Ng, K. W. (1990), *Symmetric Multivariate and Related Distributions*, London: Chapman and Hall Ltd. [3,5]

Feragen, A., Hauberg, S., Nielsen, M., and Lauze, F. (2011), "Means in Space of Tree-Like Shapes," in *2011 International Conference on Computer Vision*. [2]

Fréchet, M. (1948), "Les éléments aléatoires de nature quelconque dans un espace distancié," *Annales de l'Institut Henri Poincaré*, 10, 103–111. [4]

Gelman, A., Meng, X.-L., and Stern, H. (1996), "Posterior Predictive Assessment of Model Fitness via Realized Discrepancies," *Statistica Sinica*, 6, 733–807. [10]

Ghosal, S., and van der Vaart, A. (2017), *Fundamentals of Nonparametric Bayesian Inference*, Cambridge Series in Statistical and Probabilistic Mathematics (Vol. 44), Cambridge: Cambridge University Press. [8]

Ginestet, C. E., Li, J., Balachandran, P., Rosenberg, S., and Kolaczyk, E. D. (2017), "Hypothesis Testing for Network Data in Functional Neuroimaging," *The Annals of Applied Statistics*, 11, 725–750. [2,16]

Gollini, I., and Murphy, T. B. (2016), "Joint Modeling of Multiple Network Views," *Journal of Computational and Graphical Statistics*, 25, 246–265. [2]

Hajek, B., Wu, Y., and Xu, J. (2017), "Information Limits for Recovering a Hidden Community," *IEEE Transactions on Information Theory*, 63, 4729–4745. [15]

Hammond, D. K., Gur, Y., and Johnson, C. R. (2013), "Graph Diffusion Distance: A Difference Measure for Weighted Graphs Based on the Graph Laplacian Exponential Kernel," in *2013 Global Conference on Signal and Information Processing (GlobalSIP)*, IEEE. [3,5,10,11]

Hoff, P. D., Raftery, A. E., and Handcock, M. S. (2002), "Latent Space Approaches to Social Network Analysis," *Journal of the American Statistical Association*, 97, 1090–1098. [2]

Johnson, V. E. (2004), "A Bayesian χ^2 Test for Goodness-of-Fit," *The Annals of Statistics*, 32, 29–46. [10]

Kolaczyk, E. D., Lin, L., Rosenberg, S., and Walters, J. (2020), "Averages of Unlabeled Networks: Geometric Characterization and Asymptotic Behavior," *The Annals of Statistics*, 48, 514–538. [2]

Le, C. M., Levin, K., and Levina, E. (2018), "Estimating a Network From Multiple Noisy Realizations," *Electronic Journal of Statistics*, 12, 4697–4740. [2,14]

Lynall, M.-E., Bassett, D. S., and Kerwin, R. (2010), "Functional Connectivity and Brain Networks in Schizophrenia," *Journal of Neuroscience*, 30, 9477–9487. [1]

Mardia, K. V., and Dryden, I. L. (1998), "The Complex Watson Distribution and Shape Analysis," Technical Report, Department of Statistics, University of Leeds. [2,7]

Mézard, M., and Montanari, A. (2009), *Information, Physics, and Computation*, New York: Oxford University Press. [4]

Mitra, R., Müller, P., Liang, S., Yue, L., and Ji, Y. (2013), "A Bayesian Graphical Model for ChiP-Seq Data on Histone Modifications," *Journal of the American Statistical Association*, 108, 69–80. [2,5]

Møller, J., Pettitt, A. N., Reeves, R., and Berthelsen, K. K. (2006), "An Efficient Markov Chain Monte Carlo Method for Distributions With Intractable Normalising Constants," *Biometrika*, 93, 451–458. [7,8]

Mukherjee, S., and Speed, T. P. (2008), "Network Inference Using Informative Priors," *Proceedings of the National Academy of Sciences of the United States of America*, 105, 14313–14318. [2]

Nelson, B. G., Bassett, D. S., Camchong, J., Bullmore, E. T., and Lim, K. O. (2017), "Comparison of Large-Scale Human Brain Functional and Anatomical Networks in Schizophrenia," *NeuroImage*, 15, 439–448. [1]

Newman, M. E. J. (2018), "Network Reconstruction and Error Estimation With Noisy Network Data," arXiv no. 1703.07376v2. [2,5,14,16]

Ni, Y., Müller, P., Zhu, Y., and Ji, Y. (2018), "Heterogeneous Reciprocal Graphical Models," *Biometrics*, 74, 606–615. [2]

Nielsen, A., and Witten, D. (2018), "The Multiple Random Dot Product Graph Model," arXiv no. 1811.12172v1. [2]

Peixoto, T. P. (2018a), "Reconstructing Networks With Unknown and Heterogeneous Errors," arXiv no. 1806.07956. [14]

——— (2018b), "Network Structure From Rich But Noisy Data," *Nature, Physics*, 14, 542–545. [2,16]

Pennec, X. (2006), "Intrinsic Statistics on Riemannian Manifolds: Basic Tools for Geometric Measurements," *Journal of Mathematical Imaging and Vision*, 25, 127–154. [2]

Srivastava, A., and Klassen, E. C. (2016), *Functional and Shape Data Analysis*, New York: Springer-Verlag. [14]

Tan, L., Jasra, A., De Iorio, M., and Ebbels, T. M. D. (2017), "Bayesian Inference for Multiple Gaussian Graphical Models With Application to Metabolic Association Networks," *The Annals of Applied Statistics*, 11, 2222–2251. [2]

Tang, M., Athreya, A., Sussman, D. L., Lyzinski, V., Park, Y., and Priebe, C. (2017), "A Semiparametric Two-Sample Hypothesis Testing Problem for Random Graphs," *Journal of Computational and Graphical Statistics*, 26, 344–354. [2]

Wade, S., and Ghahramani, Z. (2018), "Bayesian Cluster Analysis: Point Estimation and Credible Balls," *Bayesian Analysis*, 13, 559–626. [2]

Wang, S., Vogelstein, J. T., and Priebe, C. E. (2017), "Joint Embedding of Graphs," arXiv no. 1703.03862. [2]

Wu, C., and Robert, C. (2017), "Average of Recentered Parallel MCMC for Big Data," arXiv no. 1706.04780v2. [14]

Zelinka, B. (1975), "On a Certain Distance Between Isomorphism Classes of Graphs," *Časopis Pro Pěstování Matematiky*, 100, 371–373. [5]

Zuo, X.-N., Anderson, J. S., Bellec, P., Birn, R. M., Biswal, B. B., Blautzik, J., Breitner, J. C., Buckner, R. L., Calhoun, V. D., Castellanos, F. X., Chen, A., Chen, B., Chen, J., Chen, X., Colcombe, S. J., Courtney, W., Craddock, R. C., Di Martino, A., Dong, H.-M., Fu, X., Gong, Q., Gorgolewski, K. J., Han, Y., He, Y., He, Y., Ho, E., Holmes, A., Hou, X.-H., Huckins, J., Jiang, T., Jiang, Y., Kelley, W., Kelly, C., King, M., LaConte, S. M., Lainhart, J. E., Lei, X., Li, H.-J., Li, K., Li, K., Lin, Q., Liu, D., Liu, J., Liu, X., Liu, Y., Lu, G., Lu, J., Luna, B., Luo, J., Lurie, D., Mao, Y., Margulies, D. S., Mayer, A. R., Meindl, T., Meyerand, M. E., Nan, W., Nielsen, J. A., O'Connor, D., Paulsen, D., Prabhakaran, V., Qi, Z., Qiu, J., Shao, C., Shehzad, Z., Tang, W., Villringer, A., Wang, H., Wang, K., Wei, D., Wei, G.-X., Weng, X.-C., Wu, X., Xu, T., Yang, N., Yang, Z., Zang, Y.-F., Zhang, L., Zhang, Q., Zhang, Z., Zhang, Z., Zhao, K., Zhen, Z., Zhou, Y., Zhu, X.-T., and Milham, M. P. (2014), "An Open Science Resource for Establishing Reliability and Reproducibility in Functional Connectomics," *Scientific Data*, 1, 140049. [2,12]