The Fréchet Mean of Inhomogeneous Random Graphs

François G. Meyer

Applied Mathematics, University of Colorado at Boulder, Boulder CO 80305
fmeyer@colorado.edu
WWW home page: https://francoismeyer.github.io

Abstract. To characterize the "average" of a sample of graphs, one can compute the sample Fréchet mean, which provides an interpretable summary of the graph sample. In this paper, we prove the following result: if we use the Hamming distance to compute distances between graphs, then the Fréchet mean of an ensemble of inhomogeneous random graph with probability matrix $\boldsymbol{P}$ is obtained by thresholding the matrix $\boldsymbol{P}$: an edge exists between the vertices i and j in the mean graph if and only if $p_{ij} > 1/2$. We prove that the result also holds for the sample mean when $\boldsymbol{P}$ is replaced with the sample mean of the adjacency matrices. This theoretical result has practical implications for random graphs ensembles such as the popular stochastic block models.

Keywords: graph metrics; Hamming distance; Karcher mean; statistical network analysis.

1 Introduction

We consider the set $\mathcal{G}$ formed by all unweighted simple labeled graphs with vertex set $\{1, \ldots, n\}$. We equip $\mathcal{G}$ with a probability measure $\mathbb{P}$. We can then characterize the geometric "average" of $\mathcal{G}$ by computing the Fréchet mean, [9], that are defined as follows.

Definition 1. *The Fréchet mean is the solution to*

$$\mathbb{E}\left[G\right] = \underset{G \in \mathcal{G}}{\text{argmin}} \sum_{H \in \mathcal{G}} d^2(G, H)\,\mathbb{P}(H) \tag{1}$$

where d is a distance defined on $\mathcal{G}$.

This notion of centrality is well adapted to metric spaces (e.g., [4,12,15] and references therein). By replacing $\mathbb{P}$ with the empirical measure, the definitions of the Fréchet mean can be extended to a sample of graphs $\left\{G^{(1)}, \ldots, G^{(N)}\right\}$, which are defined on the same vertex set $\{1, \ldots, n\}$.

Definition 2. *The sample Fréchet mean is the solution to*

$$\widehat{\mathbb{E}}_N\left[G\right] = \underset{G \in \mathcal{G}}{\text{argmin}} \frac{1}{N} \sum_{k=1}^{N} d^2(G, G^{(k)}). \tag{2}$$

Because the focus of this work is not the computation of the sample Fréchet mean, but rather a theoretical analysis of the properties that the sample Fréchet mean inherits from the probability measure $\mathbb{P}$, defined in (3), we can assume that all the graphs are defined on the same vertex set.

We note that a solution to the minimization problem (1) always exists, but need not be unique. In this work, all the results hold for any graph in the set formed by the solutions to (1). To simplify the exposition, and without any loss of generality, we therefore assume that the Fréchet mean contains a single element.

The vital role played by the Fréchet mean as a location parameter, is exemplified in the work of [16], who have created novel families of random graphs by generating random perturbations around a given Fréchet mean. In practice, the Fréchet mean itself is computed from a training set of graphs that display specific topological features of interest. To take full advantage of the training set, one needs to insure that the sample Fréchet mean inherits from the training set the desired topological structure.

1.1 Our main contributions

In this paper, we consider the probability space formed by the inhomogeneous Erdős-Rényi random graphs [2], where the probability of existence of an edge between vertices i and j is a Bernoulli random variable with parameter given by the entry (i, j) of the matrix $\boldsymbol{P}$, and all edges are independent. We use the Hamming distance to compute distances between graphs.

We address the following foundational question: what is the population – and sample – Fréchet mean of the probability $\mathcal{G}(n, \boldsymbol{P})$ in the probability space $(\mathcal{G}, \mathbb{P})$? This question is significant because inhomogeneous Erdős-Rényi random graphs provide tractable models of random graphs that are rich enough to capture many of the topological structures of real networks [2].

We show that the population mean $\mathbb{E}[G]$ can be obtained by thresholding the matrix $\boldsymbol{P}$: an edge exists between the vertices i and j in $\mathbb{E}[G]$ if and only if $p_{ij} > 1/2$. We prove that a similar result holds for the sample Fréchet mean, when $\boldsymbol{P}$ is replaced with the sample mean of the adjacency matrices.

Because sparse graphs also provide prototypical models for real networks, one would like to guarantee that the structural properties of the random graphs in $\mathcal{G}(n, \boldsymbol{P})$ are preserved when computing the Fréchet mean. Our work indicates that the Fréchet mean of an ensemble of sparse Erdős-Rényi random graphs is the empty graph, and therefore fails to capture any of the features of the graphs in the ensemble.

2 Preliminary and Notations

2.1 The inhomogeneous Erdős-Rényi random graphs

In this work, we consider inhomogeneous Erdős-Rényi random graphs [2], and we denote by $\mathcal{G}(n, \boldsymbol{P})$ [8] the probability space formed by the inhomogeneous Erdős-Rényi random graphs defined on $\{1, \dots, n\}$, where a graph G with adjacency matrix $\boldsymbol{A}$ has probability,

$$\mathbb{P}(\boldsymbol{A}) = \prod_{1 \le 1 < j \le n} \left[p_{ij}\right]^{a_{ij}} \left[1 - p_{ij}\right]^{1-a_{ij}} . \tag{3}$$

The $n \times n$ matrix of probabilities $\boldsymbol{P} = \left[p_{ij}\right]$, where $0 \le p_{ij} \le 1$ and $p_{ii} = 0$, encodes the probability of edges. We denote by $\mathcal{S}$ the set of $n \times n$ adjacency matrices of unweighted simple labeled graphs with vertex set $\{1, \dots, n\}$,

$$\mathcal{S} = \left\{\boldsymbol{A} \in \{0,1\}^{n \times n}; \text{where } a_{ij} = a_{ji}, \text{and } a_{i,i} = 0;\ 1 \le i < j \le n\right\} . \tag{4}$$

We can identify $\mathcal{G}(n, \boldsymbol{P})$ with the probability space $(\mathcal{S}, \mathbb{P})$, where $\mathbb{P}$ is defined by (3). This probability space includes stochastic block models [1], which have great practical importance.

2.2 The Hamming distance

We use the Hamming distance to compare two graphs defined on the same vertex set.

Definition 3. *Let G, G' be two graphs in $\mathcal{G}$ with adjacency matrix $\boldsymbol{A}$ and $\boldsymbol{A}'$ respectively. The Hamming distance between G and G' is given by*

$$d_H(G, G') = \sum_{1 \le i < j \le n} |a_{ij} - a'_{ij}|. \tag{5}$$

2.3 The population and sample Fréchet functions for the Hamming distance

We collect some basic facts about the Fréchet functions for the Hamming distance. We start with the expression of the Fréchet function for the probability space $(\mathcal{S}, \mathbb{P})$.

Definition 4. *We denote by F_2 the Fréchet function associated with the Fréchet mean*

$$F_2(\boldsymbol{B}) = \sum_{\boldsymbol{A}\in\mathcal{S}} d_H^2(\boldsymbol{A},\boldsymbol{B})\,\mathbb{P}(\boldsymbol{B})\,. \tag{6}$$

We provide the following expression for the Fréchet function associated to the mean.

Lemma 1. *Let $\boldsymbol{B} \in \mathcal{S}$, let $\mathcal{E}(\boldsymbol{B})$ be the set of edges of the graph associated to $\boldsymbol{B}$. Then*

$$F_2(\boldsymbol{B}) = \left[\sum_{(i,j)\in\mathcal{E}(\boldsymbol{B})} (1-2p_{ij}) + \sum_{1\le i<j\le n} p_{ij}\right]^2 + \sum_{1\le i<j\le n} p_{ij}(1-p_{ij}). \tag{7}$$

Proof of Lemma 1 *The proof is given in Appendix, section A.2.*

We now consider N independent random graphs sampled from $\mathcal{G}\left(n,\boldsymbol{P}\right)$, $\{G_k\}_{1\le k\le N}$. We denote by$\boldsymbol{A}^{(k)}$ the adjacency matrix of graph G_k. The sample Fréchet function – for the Fréchet mean defined by (2) – is given by

Definition 5. *We denote by $\widehat{F}_q$ the sample Fréchet function associated with the sample Fréchet or mean,*

$$\widehat{F}_2(\boldsymbol{B}) = \frac{1}{N}\sum_{k=1}^{N} d_H^2(\boldsymbol{A}^{(k)},\boldsymbol{B}). \tag{8}$$

We have the following expression for the sample Fréchet function associated to the sample mean, which corresponds to the sample version of (7).

Lemma 2. *Let $\boldsymbol{B} \in \mathcal{S}$, let $\mathcal{E}(\boldsymbol{B})$ be the set of edges of the graph associated to $\boldsymbol{B}$. Then*

$$\widehat{F}_2(\boldsymbol{B}) = \left[\sum_{(i,j)\in\mathcal{E}(\boldsymbol{B})} \left(1-2\widehat{\mathbb{E}}_N\left[a_{ij}\right]\right) + \sum_{1\le i<j\le n} \widehat{\mathbb{E}}_N\left[a_{ij}\right]\right]^2 + \sum_{1\le i<j\le n} \widehat{\mathbb{E}}_N\left[a_{ij}\right]\left(1-\widehat{\mathbb{E}}_N\left[a_{ij}\right]\right) \tag{9}$$

$$- \sum_{1\le i<j\le n}\ \sum_{\substack{1\le i'<j'\le n\\ (i,j)\neq(i',j')}} \left(\widehat{\mathbb{E}}_N\left[a_{ij}\right]\widehat{\mathbb{E}}_N\left[a_{i'j'}\right] - \widehat{\mathbb{E}}_N\left[\rho_{ij,i'j'}\right]\right)$$

$$+\,4 \sum_{(i,j)\in\mathcal{E}(\boldsymbol{B})}\ \sum_{(i',j')\in\overline{\mathcal{E}}(\boldsymbol{B})} \left(\widehat{\mathbb{E}}_N\left[a_{ij}\right]\widehat{\mathbb{E}}_N\left[a_{i'j'}\right] - \widehat{\mathbb{E}}_N\left[\rho_{ij,i'j'}\right]\right) \tag{10}$$

where

$$\widehat{\mathbb{E}}_N\left[a_{ij}\right] = \frac{1}{N}\sum_{k=1}^{N} a_{ij}^{(k)} \quad \text{and} \quad \widehat{\mathbb{E}}_N\left[\rho_{ij,i'j'}\right] = \frac{1}{N}\sum_{k=1}^{N} a_{ij}^{(k)}a_{i'j'}^{(k)} \tag{11}$$

Proof of Lemma 2 *The proof is given in Appendix, section A.3.*

2.4 Concentration of the sample Fréchet functions for large sample size.

In the following lemma, we show that for large sample size N, the sample Fréchet function $\widehat{F}_2$ concentrate around its population counterpart.

Lemma 3. *For all $\delta \in (0,1)$, there exists N_0, such that for all $N \ge N_0$,*

$$\widehat{F}_2(\boldsymbol{B}) = \left[\sum_{(i,j)\in\mathcal{E}(\boldsymbol{B})} (1-2p_{ij}) + \sum_{1\le i<j\le n} p_{ij}\right]^2 + \sum_{1\le i<j\le n} p_{ij}\left(1-p_{ij}\right) + O\!\left(\frac{1}{\sqrt{N}}\right), \tag{12}$$

with probability $1-\delta$ over the realization of the sample $\{G_k\}_{1\le k\le N}$.

Proof of Lemma 3 *The full detailed rigorous proof is provided in Appendix, section A.4; we give in the following the key steps.*

The analysis relies on the concentration of Bernoulli random variables in the Fréchet function. Since $a_{ij}^{(k)}$ is a Bernoulli random variable, it concentrates around its mean p_{ij} for large sample size, and therefore

$$\widehat{\mathbb{E}}_N\left[a_{ij}\right] \approx p_{ij} \quad \textit{for large } N. \tag{13}$$

Consequently the first two terms in $\widehat{F}_2(\boldsymbol{B})$ in (9), on the first row of (10), match the corresponding term in $F_2(\boldsymbol{B})$ in (7), for large N, and yield the main terms in (12).

The next two terms (second and third lines) of (10) become negligible for large sample size. Indeed, since $(i,j) \neq (i',j')$ in both terms, the random variable $a_{ij}^{(k)}a_{i'j'}^{(k)}$ is Bernoulli with parameter $p_{ij}p_{i'j'}$, which concentrates around its mean. Consequently, we have

$$\widehat{\mathbb{E}}_N\left[\rho_{ij,i'j'}\right] \approx p_{ij}p_{i'j'} \quad \textit{for large } N. \tag{14}$$

Also, since $(i,j) \neq (i',j')$, the random variables $\widehat{\mathbb{E}}_N\left[a_{ij}\right]$ and $\widehat{\mathbb{E}}_N\left[a_{i'j'}\right]$ are independent, and each concentrates around its respective mean, so

$$\widehat{\mathbb{E}}_N\left[a_{ij}\right]\widehat{\mathbb{E}}_N\left[a_{i'j'}\right] \approx p_{ij}p_{i'j'} \quad \textit{for large } N, \tag{15}$$

and thus

$$\widehat{\mathbb{E}}_N\left[\rho_{ij,i'j'}\right] - \widehat{\mathbb{E}}_N\left[a_{ij}\right]\widehat{\mathbb{E}}_N\left[a_{i'j'}\right] \approx p_{ij}p_{i'j'} - p_{ij}p_{i'j'} = 0 \quad \textit{for large } N. \tag{16}$$

3 Main Results

We first consider the (population) Fréchet mean. Our analysis is concerned with the probability measure (3), associated with the inhomogeneous Erdős-Rényi random graphs $\mathcal{G}(n, \boldsymbol{P})$.

3.1 The population Fréchet mean of graphs in $\mathcal{G}(n, \mathbb{P})$

Theorem 1. *Let $\boldsymbol{P} = \left[p_{ij}\right]$ be an $n \times n$ probability matrix. The Fréchet mean $\mathbb{E}\left[\boldsymbol{A}\right]$ of the probability measure (3), associated with the inhomogeneous Erdős-Rényi random graphs $\mathcal{G}(n, \boldsymbol{P})$ is equal to*

$$[\mathbb{E}\left[\boldsymbol{A}\right]]_{ij} = \begin{cases} 1 & \textit{if} \quad p_{ij} > 1/2, \\ 0 & \textit{otherwise.} \end{cases} \tag{17}$$

Proof. By lemma 1, we seek the matrix $\boldsymbol{B}$, with edge set $\mathcal{E}(\boldsymbol{B})$, that minimizes the function

$$\left[\sum_{(i,j)\in\mathcal{E}(\boldsymbol{B})} \left(1 - 2p_{ij}\right) + \sum_{1\leq i<j\leq n} p_{ij}\right]^2. \tag{18}$$

Let us denote by x the following expression $x \overset{\text{def}}{=} \sum_{(i,j)\in\mathcal{E}(\boldsymbol{B})}\left(1 - 2p_{ij}\right)$. We observe that since $0 \leq p_{ij} \leq 1$, we have

$$-\sum_{1\leq i<j\leq n} p_{ij} \leq -\sum_{(i,j)\in\mathcal{E}(\boldsymbol{B})} p_{ij} \leq x \leq \sum_{(i,j)\in\mathcal{E}(\boldsymbol{B})} 1 \leq n(n-1)/2. \tag{19}$$

We define, $f(x) \overset{\text{def}}{=} \left[x + \sum_{1\leq i<j\leq n} p_{ij}\right]^2$. The function $f(x)$ is equal to $F_2(\boldsymbol{B}) - \sum_{1\leq i<j\leq n} p_{ij}(1-p_{ij})$. Minimizing f is therefore equivalent to minimizing F_2. We seek x^* that minimizes $f(x)$ over the interval $\left[-\sum_{1\leq i<j\leq n} p_{ij},\ n(n-1)/2\right]$. We note immediately that x^* cannot be positive.

Otherwise, we would get $f(x^*) > f(0)$, since f is minimum at $-\sum_{1\leq i<j\leq n} p_{ij}$, and is increasing over $[-\sum_{1\leq i<j\leq n} p_{ij}, \infty)$.

Because f is convex, and has a global minimum at $-\sum_{1\leq i<j\leq n} p_{ij}$, the optimal value x^* is obtained by minimizing the distance from x^* to $-\sum_{1\leq i<j\leq n} p_{ij}$. We have

$$x - (-\sum_{1\leq i<j\leq n} p_{ij}) = \sum_{(i,j)\in\mathcal{E}(\boldsymbol{B})} (1-2p_{ij}) + \sum_{1\leq i<j\leq n} p_{ij} \geq \sum_{(i,j);1-2p_{ij}<0} (1-2p_{ij}) + \sum_{1\leq i<j\leq n} p_{ij}. \quad (20)$$

The lower bound (20) is independent of $\boldsymbol{B}$, and can be obtained by choosing,

$$[\mathbb{E}\left[\boldsymbol{A}\right]]_{ij} = \begin{cases} 1 & \text{if} \quad p_{ij} > 1/2, \\ 0 & \text{otherwise.} \end{cases} \quad (21)$$

as advertised in the theorem. □

3.2 The sample Fréchet mean of graphs in $\mathcal{G}(n, \mathbb{P})$

We now turn our attention to the sample Fréchet mean. This estimator has recently been used for the statistical analysis of graph-valued data (e.g., [6,10,16,19] and references therein). In general, the computation of the sample Fréchet mean using the Hamming distance is NP-hard [3]. For this reason, several alternatives have been proposed to the minimization problem (2), (e.g., [7,10] and references therein).

Before we present the second main result, we take a short detour through the sample Fréchet median graph [11,13,18]. The sample Fréchet median graph is the solution to the following minimization problem, wherein d_H^2 is replaced by d_H,

$$\widehat{m}_N\left[G\right] = \underset{G\in\mathcal{G}}{\operatorname{argmin}} \frac{1}{N}\sum_{k=1}^{N} d_H(G, G^{(k)}). \quad (22)$$

Because we use the Hamming distance, the sample Fréchet median graph can be characterized analytically (unlike the sample Fréchet mean graph).

Lemma 4. *The adjacency matrix $\widehat{m}_N\left[\boldsymbol{A}\right]$ of the sample median graph is given by the majority rule*

$$[\widehat{m}_N\left[\boldsymbol{A}\right]]_{ij} = \begin{cases} 0 & \text{if } \sum_{k=1}^{N} a_{ij}^{(k)} < N/2, \\ 1 & \text{otherwise.} \end{cases} \quad (23)$$

for $i, j \in \{1, \ldots, n\}$.

Proof of Lemma 4 *The result is classic, and we omit the proof, which can be found for instance in [5].*

We now come back to the second main contribution, where we prove that the sample Fréchet mean graph of N independent random graphs sampled from $\mathcal{G}\left(n, \boldsymbol{P}\right)$ is asymptotically equal (for large sample size) to the sample Fréchet median graph, with high probability.

Theorem 2. *Let $\boldsymbol{P} = \left[p_{ij}\right]$ be $n\times n$ symmetric matrix with entries $0 \leq p_{ij} \leq 1$. For all $\delta \in (0,1)$, there exists N_0, such that for all $N \geq N_0$, the sample Fréchet median $\widehat{m}_N\left[\boldsymbol{A}\right]$ and the sample Fréchet mean $\widehat{\mathbb{E}}_N\left[\boldsymbol{A}\right]$ are equal to*

$$\forall i, j \in [n], \quad \left[\widehat{m}_N\left[\boldsymbol{A}\right]\right]_{ij} = \left[\widehat{\mathbb{E}}_N\left[\boldsymbol{A}\right]\right]_{ij} = \begin{cases} 1 & \text{if} \quad p_{ij} > 1/2, \\ 0 & \text{otherwise.} \end{cases} \quad (24)$$

with probability $1-\delta$ over the realizations of the N inhomogeneous Erdős-Rényi random graphs, $\left\{G^{(1)}, \ldots, G^{(N)}\right\}$, sampled from $\mathcal{G}\left(n, \boldsymbol{P}\right)$.

Proof. Because of lemma 3, for all $\delta \in (0,1)$, there exists N_0, such that for all $N \geq N_0$,

$$\widehat{F}_2(\boldsymbol{B}) = \left[\sum_{(i,j)\in\mathcal{E}(\boldsymbol{B})} \left(1 - 2p_{ij}\right) + \sum_{1\leq i<j\leq n} p_{ij} \right]^2 + \sum_{1\leq i<j\leq n} p_{ij}\left(1 - p_{ij}\right) + O\left(\frac{1}{\sqrt{N}}\right), \tag{25}$$

with probability $1-\delta$ over the realization of the sample $\{G_k\}_{1\leq k\leq N}$. For N large enough, the main term dominates the the expression of $\widehat{F}_2(\boldsymbol{B})$, and we can neglect the $O\left(1/\sqrt{N}\right)$ term. We are left with the expression of the Fréchet function for the population mean, given by (7), in lemma 1. The minimum of $\widehat{F}_2(\boldsymbol{B})$ is achieved for the adjacency matrix given by the population Fréchet mean, $\mathbb{E}\left[\boldsymbol{A}\right]$, in (17), which corresponds to the majority rule for the sample Fréchet median graph given by (23). □

4 Simulation Studies

This section provides the results of experiments that were conducted to compare our theoretical results to numerical simulations. We first studied the concentration of the sample Fréchet function for large sample size, captured by lemma 3. We then computed the deviation of the sample Fréchet mean graph, given by theorem 2, away from the population Fréchet mean graph, given by theorem 1. Finally, we studied the phase transition between the empty graph and the complete graph, as we varied the prior on the probability matrix $\boldsymbol{P}$: as all the entries in $\boldsymbol{P}$ become smaller than $1/2$ with high probability, the sample mean graph approaches the empty graph; whereas when all the entries $\boldsymbol{P}$ become larger than $1/2$ with high probability, the sample mean graph approaches the complete graph.

All graphs were generated using the $\mathcal{G}\left(n, \boldsymbol{P}\right)$ model (3). The probability matrix $\boldsymbol{P}$ was chosen randomly using independent (up to symmetry) beta random variables, $p_{ij} \sim \text{beta}(A, B)$. The parameters A and B are specified for each set of experiments. The sample Fréchet mean was computed using the approximation provided by (24), in theorem 2. The software used to conduct the experiments is publicly available here [17].

4.1 Concentration of the sample Fréchet function

We first illustrate the concentration of the sample Fréchet functions for large sample size, described by lemma 3. Fig. 1-left displays the mean error between the population Fréchet function $F_2(\boldsymbol{B})$ given by (7), and the sample Fréchet function $\widehat{F}_2(\boldsymbol{B})$ given by (10). We vary the sample size for $N \in [10, 7079]$. For each sample size N, we generated N independent random graphs $G_1, \ldots, G_N$.

We first report the average error between $F_2(\boldsymbol{B})$ and $\widehat{F}_2(\boldsymbol{B})$, $\widehat{\mathbb{E}}_{N_B}\left[F_2(\boldsymbol{B}) - \widehat{F}_q(\boldsymbol{B})\right]$, computed using a sample of $N_B = 16$ independent random graphs $\boldsymbol{B}_1, \ldots, \boldsymbol{B}_{N_B}$, sampled from $\mathcal{G}\left(n, \boldsymbol{P}\right)$. We evaluated $F_2(\boldsymbol{B}_i)$, given by (7), and the sample Fréchet function $\widehat{F}_2(\boldsymbol{B}_i)$, for $i = 1, \ldots, N_B$. The sample average error between the population and the sample Fréchet functions was computed,

$$\widehat{\mathbb{E}}_{N_B}\left[F_2(\boldsymbol{B}) - \widehat{F}_q(\boldsymbol{B})\right] = \frac{1}{N_B}\sum_{i=1}^{N_B}\left|F_2(\boldsymbol{B}_i) - \widehat{F}_2(\boldsymbol{B})\right|, \tag{26}$$

and corresponds to a point in Fig. 1-left. We repeated this simulation 64 times to create 64 different values for each N in Fig. 1-left. A linear regression was computed and is displayed (in blue) in Fig. 1-left. The slope was found to be -0.5028, confirming the $1/\sqrt{N}$ decay of the error predicted by lemma 3.

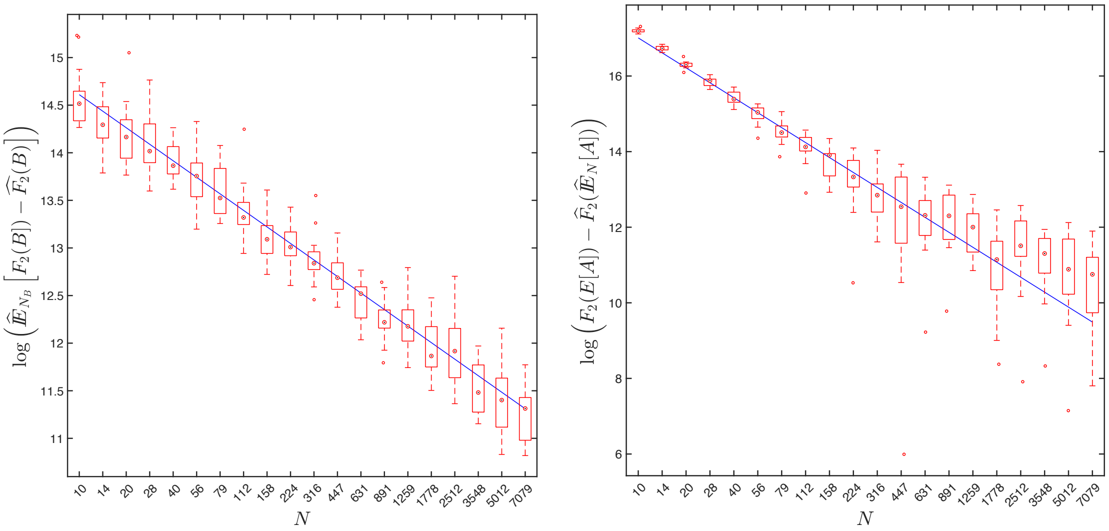

Fig. 1. Left: mean error between the population Fréchet function $F_2(\boldsymbol{B})$, given by (7), and the sample Fréchet function $\widehat{F}_2(\boldsymbol{B})$, given by (10), as a function of the sample size N. Right: error between the population Fréchet function (7) evaluated at the population Fréchet mean (17), $F_2(\mathbb{E}\left[\boldsymbol{A}\right])$, and the sample Fréchet function (10) evaluated at the sample Fréchet mean (24), $\widehat{F}_2\Big(\widehat{\mathbb{E}}_N\left[\boldsymbol{A}\right]\Big)$, as a function of the sample size N.

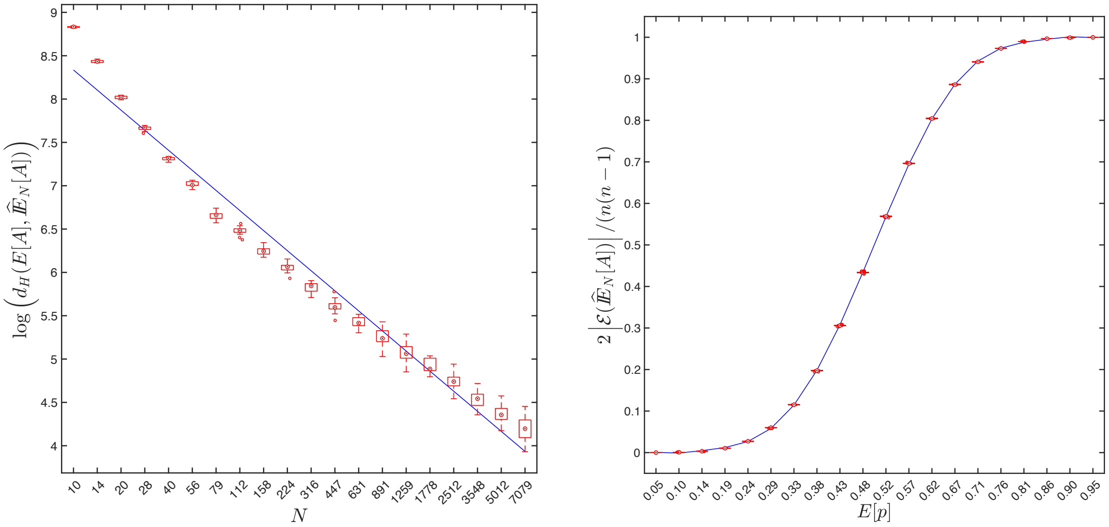

Fig. 2. Left: Hamming distance between the population Fréchet mean (17), $\mathbb{E}\left[\boldsymbol{A}\right]$, and the sample Fréchet mean (24), $\widehat{\mathbb{E}}_N\left[\boldsymbol{A}\right]$. Right: Number of edges, $\left|\mathcal{E}(\widehat{\mathbb{E}}_N\left[\boldsymbol{A}\right])\right|$, (as a fraction of the maximum number of edges) in the sample Fréchet mean graph as a function of the location parameter $\mathbb{E}\left[p_i j\right]$ of the beta distribution prior on $\boldsymbol{P}$.

We also display in Fig. 1-right the difference between the population Fréchet function evaluated at its minimum (the Fréchet mean) and the sample Fréchet function evaluated at its minimum (the sample Fréchet mean). A linear regression was evaluated and is displayed (in blue) in Fig. 1-right. The slope was found to be -1.1456. The concentration of the sample Fréchet function at its minimum occurs at a rate $1/N$, which is faster than the estimate predicted by lemma 3.

4.2 Deviation of the sample mean graph away from the population mean graph

To estimate the deviation of the sample Fréchet mean graph away from the population Fréchet mean graph, we computed the Hamming distance between the two graphs as a function of the sample size N. Fig 2-left shows the Hamming distance between the population and sample Fréchet means. A linear regression was computed and is displayed (in blue) in Fig. 2. The slope was found to be -0.6707, suggesting that the sample Fréchet mean converges toward the population Fréchet mean at a rate $1/N^{2/3}$, which is faster than the rate predicted by lemma 3.

4.3 The sharp threshold in the context of the beta prior

In this last simulation, we explored the switching of the mean graph from the empty graph to the complete graph, as the entries in the probability matrix $\boldsymbol{P}$ shift from being all less than $1/2$ to being all larger than $1/2$. To control the size of the entries in $\boldsymbol{P}$, we chose $\boldsymbol{P}$ at random using independent (up to symmetry) beta random variables,

$$p_{ij} \sim \text{beta}(A, B). \tag{27}$$

For a fixed value of A, we have $\mathbb{E}\left[p_{ij}\right] = A/(A+B)$. In our experiments, we fix $A+B$, and we let A vary over the interval $(0, A+B)$, thereby exploring the range $[0,1]$ for $\mathbb{E}\left[p_{ij}\right]$. Fig. 2-right displays the number of edges, $\left|\mathcal{E}(\widehat{\mathbb{E}}_N\left[\boldsymbol{A}\right])\right|$, (as a fraction of the maximum number of edges) in the sample Fréchet mean graph as a function of the location parameter $\mathbb{E}\left[p_{ij}\right]$ of the beta distribution prior on $\boldsymbol{P}$. The experiment confirms that the Fréchet mean of inhomogeneous Erdős-Rényi random graphs exhibit a sharp threshold: it rapidly switches from the empty graph to the complete graph as the expected edge probability $\mathbb{E}\left[p_{inj}\right]$ becomes larger than $1/2$.

5 Discussion and Conclusion

In this work we derived the expression for the population Fréchet mean for inhomogeneous Erdős-Rényi random graphs. We proved that the sample Fréchet mean was consistent, and could be estimated using a simple thresholding rule. Our results have several practical implications.

First, our work implies that the sample Fréchet mean computed from a training set of graphs, which display specific topological features of interest, will not inherit from the training set the desired topological structure. Indeed, in the context of inhomogeneous Erdős-Rényi random graphs, the (population or sample – for large sample size) the Fréchet mean graph no longer captures the edge density encoded by the edge probability.

Our answer to the question of [16]: "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?" is therefore disappointing in the context of the probability space $\mathcal{G}\left(n, \boldsymbol{P}\right)$. While the Fréchet mean is an element of $\mathcal{G}\left(n, \boldsymbol{P}\right)$, it only provides a simplistic sketch of that probability space. Consider for instance sparse graphs where $\min p_{ij} < 1/2$ (e.g., graphs with $o\left(n^2\right)$ but $\omega(n)$ edges), then the sample Fréchet mean is the empty graph.

On a more positive side, our analysis provides a theoretical justification for several algorithms designed to recover a graph from noisy measurements of its adjacency matrix. For instance, the authors in [14] devise a method to recover a fixed network from unlabeled noisy samples. Instead of estimating the Fréchet mean, they compute the sample mean of the noisy adjacency matrices, and threshold the sample mean to recover an unweighted graph. Our results offer a theoretical justification of their approach, if one assumes that the noisy graphs are aligned and are realizations of inhomogeneous Erdős-Rényi random graphs, with an unknown edge probability matrix $\boldsymbol{P}$.

Acknowledgments

F.G.M was supported by the National Natural Science Foundation (CCF/CIF 1815971).

References

1. Abbe, E.: Community detection and stochastic block models: recent developments. The Journal of Machine Learning Research 18(1), 6446–6531 (2017)
2. Bollobás, B., Janson, S., Riordan, O.: The phase transition in inhomogeneous random graphs. Random Structures & Algorithms 31(1), 3–122 (2007)
3. Chen, J., Hermelin, D., Sorge, M.: On Computing Centroids According to the p-Norms of Hamming Distance Vectors. In: 27th Annual European Symposium on Algorithms (ESA 2019). vol. 144, pp. 28:1–28:16. Dagstuhl, Germany (2019)
4. Chowdhury, S., Mémoli, F.: The metric space of networks (2018)
5. Devroye, L., Györfi, L., Lugosi, G.: A probabilistic theory of pattern recognition, vol. 31. Springer Science & Business Media (2013)
6. Dubey, P., Müller, H.G.: Fréchet change-point detection. The Annals of Statistics 48(6), 3312–3335 (2020)
7. Ferrer, M., Valveny, E., Serratosa, F., Riesen, K., Bunke, H.: Generalized median graph computation by means of graph embedding in vector spaces. Pattern Recognition 43(4), 1642–1655 (2010)
8. Frieze, A., Karoński, M.: Introduction to random graphs. Cambridge University Press (2016)
9. Fréchet, M.: Les espaces abstraits et leur utilité en statistique théorique et même en statistique appliquée. Journal de la Société Française de Statistique 88, 410–421 (1947)
10. Ginestet, C.E., Li, J., Balachandran, P., Rosenberg, S., Kolaczyk, E.D.: Hypothesis testing for network data in functional neuroimaging. The Annals of Applied Statistics 11(2), 725–750 (2017)
11. Han, F., Han, X., Liu, H., Caffo, B., et al.: Sparse median graphs estimation in a high-dimensional semiparametric model. The Annals of Applied Statistics 10(3), 1397–1426 (2016)
12. Jain, B.J.: Statistical graph space analysis. Pattern Recognition 60, 802–812 (2016)
13. Jiang, X., Munger, A., Bunke, H.: On median graphs: properties, algorithms, and applications. IEEE Transactions on Pattern Analysis and Machine Intelligence 23(10), 1144–1151 (2001)
14. Josephs, N., Li, W., Kolaczyk, E.D.: Network recovery from unlabeled noisy samples (2021)
15. Kolaczyk, E.D., Lin, L., Rosenberg, S., Walters, J., Xu, J., et al.: Averages of unlabeled networks: Geometric characterization and asymptotic behavior. The Annals of Statistics 48(1), 514–538 (2020)
16. Lunagómez, S., Olhede, S.C., Wolfe, P.J.: Modeling network populations via graph distances. Journal of the American Statistical Association pp. 1–18 (2020)
17. Meyer, F.G.: The Fréchet Mean of Inhomogeneous Random Graphs. https://francoismeyer.github.io/frechet-mean (2021)
18. Mukherjee, L., Singh, V., Peng, J., Xu, J., Zeitz, M.J., Berezney, R.: Generalized median graphs and applications. Journal of Combinatorial Optimization 17(1), 21–44 (2009)
19. Zambon, D., Alippi, C., Livi, L.: Change-point methods on a sequence of graphs. IEEE Transactions on Signal Processing 67(24), 6327–6341 (2019)

A Proofs of the main results

A.1 A simpler expression for the Hamming distance squared

We introduce a small technical lemma that provides a simpler expression for the Hamming distance squared, which will simplify many of the proofs.

Lemma 5. *Let $\boldsymbol{A}$ and $\boldsymbol{B}$ two matrices in $\mathcal{S}$. Let $\mathcal{E}(\boldsymbol{B})$ be the set of edges of $\boldsymbol{B}$, and $\overline{\mathcal{E}}(\boldsymbol{B})$ be the set of nonedges of $\boldsymbol{B}$. Then,*

$$d_H^2(\boldsymbol{A},\boldsymbol{B}) = \left[\sum_{1\le i<j\le n} a_{ij}\right]^2 + |\mathcal{E}(\boldsymbol{B})|^2 + 2|\mathcal{E}(\boldsymbol{B})|\left[\sum_{(i,j)\in\overline{\mathcal{E}}(\boldsymbol{B})} a_{ij} - \sum_{(i,j)\in\mathcal{E}(\boldsymbol{B})} a_{ij}\right] - 4\sum_{(i,j)\in\mathcal{E}(\boldsymbol{B})}\ \sum_{(i',j')\in\overline{\mathcal{E}}(\boldsymbol{B})} a_{ij}a_{i'j'} \tag{28}$$

Proof of Lemma 5 *We denote by $\mathcal{E}(\boldsymbol{A})$ the set of edges of $\boldsymbol{A}$, and by $\mathcal{E}(\boldsymbol{B})$ the set of edges of $\boldsymbol{B}$. We observe that*

$$\begin{aligned} d_H(\boldsymbol{A},\boldsymbol{B}) &= \sum_{1\le i<j\le n}\left\{a_{ij}+b_{ij}-2a_{ij}b_{ij}\right\} = \sum_{1\le i<j\le n} a_{ij} + |\mathcal{E}(\boldsymbol{B})| - 2\sum_{1\le i<j\le n} a_{ij}b_{ij} \\ &= \sum_{1\le i<j\le n} a_{ij} + |\mathcal{E}(\boldsymbol{B})| - 2\sum_{(i,j)\in\mathcal{E}(\boldsymbol{B})} a_{ij}, \end{aligned} \tag{29}$$

Taking the square of (29), we get

$$
\begin{aligned}
d_H^2(\boldsymbol{A},\boldsymbol{B}) = &\left[\sum_{1\le i<j\le n} a_{ij}\right]^2 + |\mathcal{E}(\boldsymbol{B})|^2 + 4\left[\sum_{(i,j)\in\mathcal{E}(\boldsymbol{B})} a_{ij}\right]^2 \\
&-4\left[\sum_{1\le i<j\le n} a_{ij}\right]\left[\sum_{(i,j)\in\mathcal{E}(\boldsymbol{B})} a_{ij}\right] - 4|\mathcal{E}(\boldsymbol{B})|\left[\sum_{(i,j)\in\mathcal{E}(\boldsymbol{B})} a_{ij}\right] + 2|\mathcal{E}(\boldsymbol{B})|\left[\sum_{1\le i<j\le n} a_{ij}\right]
\end{aligned}
$$

We denote by $\overline{\mathcal{E}}(\boldsymbol{B})$ the set of "nonedges" in $\boldsymbol{B}$, that is

$$
\overline{\mathcal{E}}(\boldsymbol{B}) = \left\{(i,j); 1\le i<j\le n, b_{ij}=0\right\}. \tag{30}
$$

We can then break the sum $\sum_{1\le i<j\le n} a_{ij}$ into two parts: a sum over edges in $\boldsymbol{B}$, $\mathcal{E}(\boldsymbol{B})$, and a sum over nonedges in $\boldsymbol{B}$, $\overline{\mathcal{E}}(\boldsymbol{B})$, which leads to

$$
\begin{aligned}
d_H^2(\boldsymbol{A},\boldsymbol{B}) = &\left[\sum_{1\le i<j\le n} a_{ij}\right]^2 + |\mathcal{E}(\boldsymbol{B})|^2 + 4\left[\sum_{(i,j)\in\mathcal{E}(\boldsymbol{B})} a_{ij}\right]^2 - 4|\mathcal{E}(\boldsymbol{B})|\left[\sum_{(i,j)\in\mathcal{E}(\boldsymbol{B})} a_{ij}\right] \\
&-4\left[\sum_{(i,j)\in\overline{\mathcal{E}}(\boldsymbol{B})} a_{ij} + \sum_{(i,j)\in\mathcal{E}(\boldsymbol{B})} a_{ij}\right]\left[\sum_{(i,j)\in\mathcal{E}(\boldsymbol{B})} a_{ij}\right] + 2|\mathcal{E}(\boldsymbol{B})|\left[\sum_{(i,j)\in\overline{\mathcal{E}}(\boldsymbol{B})} a_{ij} + \sum_{(i,j)\in\mathcal{E}(\boldsymbol{B})} a_{ij}\right]
\end{aligned}
$$

Expanding and regrouping the terms yields

$$
\begin{aligned}
d_H^2(\boldsymbol{A},\boldsymbol{B}) = &\left[\sum_{1\le i<j\le n} a_{ij}\right]^2 + |\mathcal{E}(\boldsymbol{B})|^2 + 4\left[\sum_{(i,j)\in\mathcal{E}(\boldsymbol{B})} a_{ij}\right]^2 - 4\left[\sum_{(i,j)\in\mathcal{E}(\boldsymbol{B})} a_{ij}\right]^2 - 4\left[\sum_{(i,j)\in\overline{\mathcal{E}}(\boldsymbol{B})} a_{ij}\right]\left[\sum_{(i,j)\in\mathcal{E}(\boldsymbol{B})} a_{ij}\right] \\
&- 4|\mathcal{E}(\boldsymbol{B})|\left[\sum_{(i,j)\in\mathcal{E}(\boldsymbol{B})} a_{ij}\right] + 2|\mathcal{E}(\boldsymbol{B})|\left[\sum_{(i,j)\in\mathcal{E}(\boldsymbol{B})} a_{ij}\right] + 2|\mathcal{E}(\boldsymbol{B})|\left[\sum_{(i,j)\in\overline{\mathcal{E}}(\boldsymbol{B})} a_{ij}\right]
\end{aligned} \tag{31}
$$

and we conclude that

$$
d_H^2(\boldsymbol{A},\boldsymbol{B}) = \left[\sum_{1\le i<j\le n} a_{ij}\right]^2 + |\mathcal{E}(\boldsymbol{B})|^2 - 4\sum_{(i,j)\in\mathcal{E}(\boldsymbol{B})}\ \sum_{(i',j')\in\overline{\mathcal{E}}(\boldsymbol{B})} a_{ij}a_{i'j'} + 2|\mathcal{E}(\boldsymbol{B})|\left[\sum_{(i,j)\in\overline{\mathcal{E}}(\boldsymbol{B})} a_{ij} - \sum_{(i,j)\in\mathcal{E}(\boldsymbol{B})} a_{ij}\right].
$$

□

A.2 Proof of Lemma 1

We use lemma 5, which gives a simpler expression for the Hamming distance, and take the expectation with respect to the probability measure $\mathbb{P}$, on both sides of equation (28)

$$
\begin{aligned}
F_2(\boldsymbol{B}) &= \sum_{\boldsymbol{A}\in\mathcal{S}} d_H^2(\boldsymbol{A},\boldsymbol{B})\,\mathbb{P}(\boldsymbol{A}) \\
&= \sum_{\boldsymbol{A}\in\mathcal{S}}\left[\sum_{1\le i<j\le n} a_{ij}\right]^2\mathbb{P}(\boldsymbol{A}) + |\mathcal{E}(\boldsymbol{B})|^2 - 4\sum_{(i,j)\in\mathcal{E}(\boldsymbol{B})}\ \sum_{(i',j')\in\overline{\mathcal{E}}(\boldsymbol{B})}\ \sum_{\boldsymbol{A}\in\mathcal{S}} a_{ij}a_{i'j'}\,\mathbb{P}(\boldsymbol{A}) && (32)\\
&\quad + 2|\mathcal{E}(\boldsymbol{B})|\left[\sum_{(i,j)\in\overline{\mathcal{E}}(\boldsymbol{B})}\ \sum_{\boldsymbol{A}\in\mathcal{S}} a_{ij}\,\mathbb{P}(\boldsymbol{A}) - \sum_{(i,j)\in\mathcal{E}(\boldsymbol{B})}\ \sum_{\boldsymbol{A}\in\mathcal{S}} a_{ij}\,\mathbb{P}(\boldsymbol{A})\right]. && (33)
\end{aligned}
$$

The first term in the right-hand side of (32) can be evaluated to give

$$\begin{aligned}\sum_{A\in\mathcal{S}}\left[\sum_{1\le i<j\le n} a_{ij}\right]^2 \mathbb{P}(\boldsymbol{A}) &= \sum_{A\in\mathcal{S}}\left[\sum_{1\le i<j\le n} a_{ij}\right]\left[\sum_{1\le i'<j'\le n} a_{i'j'}\right]\mathbb{P}(\boldsymbol{A})\\ &= \sum_{A\in\mathcal{S}}\sum_{1\le i<j\le n}\sum_{1\le i'<j'\le n} a_{ij}\, a_{i'j'}\,\mathbb{P}(\boldsymbol{A}) = \sum_{1\le i<j\le n}\sum_{1\le i'<j'\le n}\sum_{A\in\mathcal{S}} a_{ij}\, a_{i'j'}\,\mathbb{P}(\boldsymbol{A})\,, \qquad (34)\end{aligned}$$

where

$$\sum_{A\in\mathcal{S}} a_{ij}a_{i'j'}\,\mathbb{P}(\boldsymbol{A}) = \mathbb{E}\left[a_{ij}a_{i'j'}\right] = \begin{cases}\mathbb{E}\left[a_{ij}\right]\mathbb{E}\left[a_{i'j'}\right] = p_{ij}^2 & \text{if } (i,j)\neq(i',j'),\\ \mathbb{E}\left[a_{ij}^2\right] = \mathbb{E}\left[a_{ij}\right] = p_{ij} & \text{if } (i,j)=(i',j').\end{cases} \qquad (35)$$

We conclude that

$$\sum_{A\in\mathcal{S}}\left[\sum_{1\le i<j\le n} a_{ij}\right]^2\mathbb{P}(\boldsymbol{A}) = \sum_{1\le i<j\le n}\sum_{\substack{1\le i'<j'\le n\\(i,j)\neq(i',j')}} p_{ij}^2 + \sum_{1\le i<j\le n} p_{ij} = \left[\sum_{1\le i<j\le n} p_{ij}\right]^2 + \sum_{1\le i<j\le n} p_{ij}(1-p_{ij}). \qquad (36)$$

Now, $\sum_{A\in\mathcal{S}} a_{ij}\,\mathbb{P}(\boldsymbol{A}) = \mathbb{E}\left[a_{ij}\right] = p_{ij}$, and because the edges $(i,j)\in\mathcal{E}(\boldsymbol{B})$ and $(i',j')\in\overline{\mathcal{E}}(\boldsymbol{B})$ are independent,

$$\mathbb{E}\left[a_{ij}a_{i'j'}\right] = \sum_{A\in\mathcal{S}} a_{ij}a_{i'j'}\,\mathbb{P}(\boldsymbol{A}) = p_{ij}p_{i'j'}. \qquad (37)$$

Therefore the second term on the right-hand side of (32) becomes

$$-4\sum_{(i,j)\in\mathcal{E}(\boldsymbol{B})}\sum_{(i',j')\in\overline{\mathcal{E}}(\boldsymbol{B})}\sum_{A\in\mathcal{S}} a_{ij}a_{i'j'}\,\mathbb{P}(\boldsymbol{A}) = -4\sum_{(i,j)\in\mathcal{E}(\boldsymbol{B})}\sum_{(i',j')\in\overline{\mathcal{E}}(\boldsymbol{B})} p_{ij}p_{i'j'} = -4\left[\sum_{(i,j)\in\mathcal{E}(\boldsymbol{B})} p_{ij}\right]\left[\sum_{(i',j')\in\overline{\mathcal{E}}(\boldsymbol{B})} p_{i'j'}\right] \qquad (38)$$

Also (33) is equal to,

$$2|\mathcal{E}(\boldsymbol{B})|\left[\sum_{(i,j)\in\overline{\mathcal{E}}(\boldsymbol{B})}\sum_{A\in\mathcal{S}} a_{ij}\,\mathbb{P}(\boldsymbol{A}) - \sum_{(i,j)\in\mathcal{E}(\boldsymbol{B})}\sum_{A\in\mathcal{S}} a_{ij}\,\mathbb{P}(\boldsymbol{A})\right] = 2|\mathcal{E}(\boldsymbol{B})|\left[\sum_{(i,j)\in\overline{\mathcal{E}}(\boldsymbol{B})} p_{ij}\right] - 2|\mathcal{E}(\boldsymbol{B})|\left[\sum_{(i,j)\in\mathcal{E}(\boldsymbol{B})} p_{ij}\right] \qquad (39)$$

We can substitute (36), (38), and (39), into (32), and (33) respectively, and we get

$$\begin{aligned}F_2(\boldsymbol{B}) &= \left[\sum_{1\le i<j\le n} p_{ij}\right]^2 + \sum_{1\le i<j\le n} p_{ij}(1-p_{ij}) + |\mathcal{E}(\boldsymbol{B})|^2 - \left[2\sum_{(i,j)\in\mathcal{E}(\boldsymbol{B})} p_{ij}\right]\left[2\sum_{(i',j')\in\overline{\mathcal{E}}(\boldsymbol{B})} p_{i'j'}\right]\\ &\quad + |\mathcal{E}(\boldsymbol{B})|\left[2\sum_{(i,j)\in\overline{\mathcal{E}}(\boldsymbol{B})} p_{ij}\right] - |\mathcal{E}(\boldsymbol{B})|\left[2\sum_{(i,j)\in\mathcal{E}(\boldsymbol{B})} p_{ij}\right]\\ &= \left[\sum_{1\le i<j\le n} p_{ij}\right]^2 + \sum_{1\le i<j\le n} p_{ij}(1-p_{ij}) + \left[|\mathcal{E}(\boldsymbol{B})| + 2\sum_{(i,j)\in\overline{\mathcal{E}}(\boldsymbol{B})} p_{ij}\right]\left[|\mathcal{E}(\boldsymbol{B})| - 2\sum_{(i,j)\in\mathcal{E}(\boldsymbol{B})} p_{ij}\right]. \qquad (40)\end{aligned}$$

We note that

$$|\mathcal{E}(\boldsymbol{B})| - 2\sum_{(i,j)\in\mathcal{E}(\boldsymbol{B})} p_{ij} = \sum_{(i,j)\in\mathcal{E}(\boldsymbol{B})} (1-2p_{ij}). \qquad (41)$$

Also,

$$|\mathcal{E}(\boldsymbol{B})| + 2 \sum_{(i,j)\in\overline{\mathcal{E}}(\boldsymbol{B})} p_{ij} = \sum_{(i,j)\in\mathcal{E}(\boldsymbol{B})} (1-2p_{ij}) + 2 \sum_{1\le i<j\le n} p_{ij}. \tag{42}$$

We can finally substitute (41) and (42) into (40), and we get

$$\begin{aligned} F_2(\boldsymbol{B}) &= \left[\sum_{1\le i<j\le n} p_{ij}\right]^2 + 2\left[\sum_{1\le i<j\le n} p_{ij}\right]\left[\sum_{(i,j)\in\mathcal{E}(\boldsymbol{B})} (1-2p_{ij})\right] + \left[\sum_{(i,j)\in\mathcal{E}(\boldsymbol{B})} (1-2p_{ij})\right]^2 + \sum_{1\le i<j\le n} p_{ij}(1-p_{ij}) \\ &= \left[\sum_{(i,j)\in\mathcal{E}(\boldsymbol{B})} (1-2p_{ij}) + \sum_{1\le i<j\le n} p_{ij}\right]^2 + \sum_{1\le i<j\le n} p_{ij}(1-p_{ij}), \end{aligned} \tag{43}$$

which matches the expression (7). □

A.3 Proof of lemma 2

The proof is a direct application of lemma 5. For each graph k, we apply equation (28), and we sum over all the graphs in the sample and divide by N to get

$$\begin{aligned} \widehat{F}_2(\boldsymbol{B}) = &|\mathcal{E}(\boldsymbol{B})|^2 + 2|\mathcal{E}(\boldsymbol{B})| \left[\sum_{(i,j)\in\overline{\mathcal{E}}(\boldsymbol{B})} \frac{1}{N}\sum_{k=1}^{N} a_{ij}^{(k)} - \sum_{(i,j)\in\mathcal{E}(\boldsymbol{B})} \frac{1}{N}\sum_{k=1}^{N} a_{ij}^{(k)}\right] \\ &+ \frac{1}{N}\sum_{k=1}^{N}\left[\sum_{1\le i<j\le n} a_{ij}^{(k)}\right]^2 - 4\sum_{(i,j)\in\overline{\mathcal{E}}(\boldsymbol{B})}\ \sum_{(i',j')\in\mathcal{E}(\boldsymbol{B})} \left[\frac{1}{N}\sum_{k=1}^{N} a_{ij}^{(k)} a_{i'j'}^{(k)}\right] \end{aligned} \tag{44}$$

Now let us denote the sample mean by

$$\widehat{\mathbb{E}}_N\left[a_{ij}\right] = \frac{1}{N}\sum_{k=1}^{N} a_{ij}^{(k)} \tag{45}$$

and the sample correlation by

$$\widehat{\mathbb{E}}_N\left[\rho_{ij,i'j'}\right] = \frac{1}{N}\sum_{k=1}^{N} a_{ij}^{(k)} a_{i'j'}^{(k)} \tag{46}$$

Then we have

$$\begin{aligned} \widehat{F}_2(\boldsymbol{B}) = &|\mathcal{E}(\boldsymbol{B})|^2 + 2|\mathcal{E}(\boldsymbol{B})| \left[\sum_{(i,j)\in\overline{\mathcal{E}}(\boldsymbol{B})} \widehat{\mathbb{E}}_N\left[a_{ij}\right] - \sum_{(i,j)\in\mathcal{E}(\boldsymbol{B})} \widehat{\mathbb{E}}_N\left[a_{ij}\right]\right] + \frac{1}{N}\sum_{k=1}^{N}\left[\sum_{1\le i<j\le n} a_{ij}^{(k)}\right]^2 \\ &-4\sum_{(i,j)\in\overline{\mathcal{E}}(\boldsymbol{B})}\ \sum_{(i',j')\in\mathcal{E}(\boldsymbol{B})} \left[\widehat{\mathbb{E}}_N\left[\rho_{ij,i'j'}\right]\right] \end{aligned} \tag{47}$$

We note that

$$\frac{1}{N}\sum_{k=1}^{N}\left[\sum_{1\le i<j\le n} a_{ij}^{(k)}\right]^2 = \sum_{1\le i<j\le n}\ \sum_{1\le i'<j'\le n} \frac{1}{N}\sum_{k=1}^{N} a_{ij}^{(k)} a_{i'j'}^{(k)} = \sum_{1\le i<j\le n}\ \sum_{1\le i'<j'\le n} \widehat{\mathbb{E}}_N\left[\rho_{ij,i'j'}\right] \tag{48}$$

Also, we have

$$\begin{aligned}
&|\mathcal{E}(\boldsymbol{B})|^2 + 2|\mathcal{E}(\boldsymbol{B})|\left[\sum_{(i,j)\in\overline{\mathcal{E}}(\boldsymbol{B})} \widehat{\mathbb{E}}_N\left[a_{ij}\right] - \sum_{(i,j)\in\mathcal{E}(\boldsymbol{B})} \widehat{\mathbb{E}}_N\left[a_{ij}\right]\right] \\
&\qquad = \left[|\mathcal{E}(\boldsymbol{B})| - 2\sum_{(i,j)\in\mathcal{E}(\boldsymbol{B})} \widehat{\mathbb{E}}_N\left[a_{ij}\right]\right]\left[|\mathcal{E}(\boldsymbol{B})| + 2\sum_{(i,j)\in\overline{\mathcal{E}}(\boldsymbol{B})} \widehat{\mathbb{E}}_N\left[a_{ij}\right]\right] \\
&\qquad\quad + 4\sum_{(i,j)\in\overline{\mathcal{E}}(\boldsymbol{B})}\ \sum_{(i',j')\in\mathcal{E}(\boldsymbol{B})} \widehat{\mathbb{E}}_N\left[a_{ij}\right]\widehat{\mathbb{E}}_N\left[a_{i'j'}\right], \\
&= \left[|\mathcal{E}(\boldsymbol{B})| - 2\sum_{(i,j)\in\mathcal{E}(\boldsymbol{B})} \widehat{\mathbb{E}}_N\left[a_{ij}\right]\right]\left[|\mathcal{E}(\boldsymbol{B})| - 2\sum_{(i,j)\in\mathcal{E}(\boldsymbol{B})} \widehat{\mathbb{E}}_N\left[a_{ij}\right] + 2\sum_{1\le i<j\le n}\ \sum_{1\le i'<j'\le n} \widehat{\mathbb{E}}_N\left[a_{ij}\right]\right] \\
&\quad + 4\sum_{(i,j)\in\overline{\mathcal{E}}(\boldsymbol{B})}\ \sum_{(i',j')\in\mathcal{E}(\boldsymbol{B})} \widehat{\mathbb{E}}_N\left[a_{ij}\right]\widehat{\mathbb{E}}_N\left[a_{i'j'}\right] \\
&= \left[|\mathcal{E}(\boldsymbol{B})| - 2\sum_{(i,j)\in\mathcal{E}(\boldsymbol{B})} \widehat{\mathbb{E}}_N\left[a_{ij}\right]\right]^2 - \sum_{1\le i<j\le n}\ \sum_{1\le i'<j'\le n} \widehat{\mathbb{E}}_N\left[a_{ij}\right]\widehat{\mathbb{E}}_N\left[a_{i'j'}\right] \\
&\quad + 4\sum_{(i,j)\in\overline{\mathcal{E}}(\boldsymbol{B})}\ \sum_{(i',j')\in\mathcal{E}(\boldsymbol{B})} \widehat{\mathbb{E}}_N\left[a_{ij}\right]\widehat{\mathbb{E}}_N\left[a_{i'j'}\right]
\end{aligned} \tag{49}$$

We can then substitute (48) and (49) into (47), and we get

$$\begin{aligned}
\widehat{F}_2(\boldsymbol{B}) &= \left[|\mathcal{E}(\boldsymbol{B})| - 2\sum_{(i,j)\in\mathcal{E}(\boldsymbol{B})} \widehat{\mathbb{E}}_N\left[a_{ij}\right]\right]^2 - \sum_{1\le i<j\le n}\ \sum_{1\le i'<j'\le n}\left[\widehat{\mathbb{E}}_N\left[a_{ij}\right]\widehat{\mathbb{E}}_N\left[a_{i'j'}\right] - \widehat{\mathbb{E}}_N\left[\rho_{ij,i'j'}\right]\right] \\
&\quad + 4\sum_{(i,j)\in\overline{\mathcal{E}}(\boldsymbol{B})}\ \sum_{(i',j')\in\mathcal{E}(\boldsymbol{B})}\left[\widehat{\mathbb{E}}_N\left[a_{ij}\right]\widehat{\mathbb{E}}_N\left[a_{i'j'}\right] - \widehat{\mathbb{E}}_N\left[\rho_{ij,i'j'}\right]\right]
\end{aligned} \tag{50}$$

Finally, we can extract from the sample correlation the term that corresponds to $(i,j) = (i',j')$,

$$\begin{aligned}
&\sum_{1\le i<j\le n}\ \sum_{1\le i'<j'\le n} \widehat{\mathbb{E}}_N\left[a_{ij}\right]\widehat{\mathbb{E}}_N\left[a_{i'j'}\right] - \widehat{\mathbb{E}}_N\left[\rho_{ij,i'j'}\right] \\
&= \sum_{1\le i<j\le n}\ \sum_{\substack{1\le i'<j'\le n\\ (i',j')\neq(i,j)}} \widehat{\mathbb{E}}_N\left[a_{ij}\right]\widehat{\mathbb{E}}_N\left[a_{i'j'}\right] - \widehat{\mathbb{E}}_N\left[\rho_{ij,i'j'}\right] + \sum_{1\le i<j\le n} \widehat{\mathbb{E}}_N\left[a_{a_{ij}}\right]\widehat{\mathbb{E}}_N\left[a_{ij}\right] - \widehat{\mathbb{E}}_N\left[\rho_{ij,ij}\right],
\end{aligned} \tag{51}$$

Now if $(i,j) = (i',j')$ we have

$$\widehat{\mathbb{E}}_N\left[\rho_{ij,ij}\right] = \sum_{k=1}^{N} a_{ij}^{(k)} a_{ij}^{(k)} = \sum_{k=1}^{N} a_{ij}^{(k)} = \widehat{\mathbb{E}}_N\left[a_{ij}\right], \tag{52}$$

and therefore

$$\begin{aligned}
&\sum_{1\le i<j\le n}\ \sum_{1\le i'<j'\le n}\left[\widehat{\mathbb{E}}_N\left[a_{ij}\right]\widehat{\mathbb{E}}_N\left[a_{i'j'}\right] - \widehat{\mathbb{E}}_N\left[\rho_{ij,i'j'}\right]\right] \\
&= \sum_{1\le i<j\le n}\ \sum_{\substack{1\le i'<j'\le n\\ (i',j')\neq(i,j)}} \widehat{\mathbb{E}}_N\left[a_{ij}\right]\widehat{\mathbb{E}}_N\left[a_{i'j'}\right] - \widehat{\mathbb{E}}_N\left[\rho_{ij,i'j'}\right] + \sum_{1\le i<j\le n} \widehat{\mathbb{E}}_N\left[a_{ij}\right]\left(\widehat{\mathbb{E}}_N\left[a_{ij}\right] - 1\right)
\end{aligned} \tag{53}$$

Substituting (53) into (50), we obtain the result advertised in the lemma. □

A.4 Proof of lemma 3

The sample mean $\widehat{\mathbb{E}}_N\left[a_{ij}\right]$ is the sum of Bernoulli random variables, and it concentrate around its mean p_{ij}. We use Hoeffding inequality to bound the variation of $\widehat{\mathbb{E}}_N\left[a_{ij}\right]$ around its mean. For each $1 \leq i < j \leq n$, we have,

$$\mathbb{P}\left(\boldsymbol{A}^{(k)} \sim \mathcal{G}(n, \boldsymbol{P})\,; \left|\widehat{\mathbb{E}}_N\left[a_{ij}\right] - p_{ij}\right| \geq \varepsilon\right) \leq \exp\left(-2N\varepsilon^2\right). \tag{54}$$

To control $\sum_{k=1}^{N} a_{ij}^{(k)}$ for all $1 \leq i < j < n$, we use a union bound, and we get,

$$\mathbb{P}\left(\forall\, 1 \leq i < j \leq n; \quad \left|\widehat{\mathbb{E}}_N\left[a_{ij}\right] - p_{ij}\right| \geq \varepsilon\right) \leq \frac{n(n-1)}{2} \exp\left(-2N\varepsilon^2\right). \tag{55}$$

We define

$$\delta = \frac{n^2}{2} \exp\left(-2N\varepsilon^2\right), \tag{56}$$

and thus

$$\varepsilon = \frac{\alpha}{\sqrt{N}} \quad \text{with} \quad \alpha = \sqrt{\log \frac{n}{\sqrt{2\delta}}}. \tag{57}$$

In summary, we have

$$\forall 1 \leq i < j < n, \quad \left|\widehat{\mathbb{E}}_N\left[a_{ij}\right] - p_{ij}\right| \leq \frac{\alpha}{\sqrt{N}}, \tag{58}$$

We now study the concentration of the sample correlation,

$$\widehat{\mathbb{E}}_N\left[\rho_{ij,i'j'}\right] = \frac{1}{N} \sum_{k=1}^{N} a_{ij}^{(k)} a_{i'j'}^{(k)}, \tag{59}$$

when the pair of edges are distinct, that is $(i, j) \neq (i', j')$. Because $(i, j) \neq (i', j')$, the terms $a_{ij}^{(k)}$ and $a_{i'j'}^{(k)}$ are always independent, and the product $a_{ij}^{(k)} a_{i'j'}^{(k)}$ is a Bernoulli random variable with parameter $p_{ij}p_{i'j'}$. We conclude that the sample correlation is the sum of Bernoulli random variables, and thus it concentrates around its mean. We use Hoeffding inequality to bound the variation of $\widehat{\mathbb{E}}_N\left[\rho_{ij,i'j'}\right]$ around its mean, $p_{ij}p_{i'j'}$.

Replicating the argument used for $\widehat{\mathbb{E}}_N\left[a_{ij}\right]$ in the paragraph above mutatis mutandis, yields

$$\mathbb{P}\left(\forall\, 1 \leq i < j \leq n, \forall\, 1 \leq i' < j' \leq n, \quad \left|\widehat{\mathbb{E}}_N\left[\rho_{ij,i'j'}\right] - p_{ij}p_{i'j'}\right| \geq \frac{\beta}{\sqrt{N}}\right) \leq \delta \tag{60}$$

with

$$\beta = \sqrt{\log \frac{n^2}{2\sqrt{\delta}}}. \tag{61}$$

In summary, we have

$$\begin{aligned} &\forall\, 1 \leq i < j \leq n, \forall\, 1 \leq i' < j' \leq n, \quad \text{with} \quad (i, j) \neq (i', j'), \\ &\qquad \widehat{\mathbb{E}}_N\left[a_{ij}\right] = p_{ij} + O\left(\frac{1}{\sqrt{N}}\right), \quad \text{and} \quad \widehat{\mathbb{E}}_N\left[\rho_{ij,i'j'}\right] = p_{i'j'} + O\left(\frac{1}{\sqrt{N}}\right). \end{aligned} \tag{62}$$

with probability $1 - 2\delta$.

We recall the expression of the sample Fréchet function for the sample mean. From lemma 2, we have

$$\begin{aligned}\widehat{F}_2(\boldsymbol{B}) = &\left[\sum_{(i,j)\in\mathcal{E}(\boldsymbol{B})}\left(1-2\widehat{\mathbb{E}}_N\left[a_{ij}\right]\right)+\sum_{1\le i<j\le n}\widehat{\mathbb{E}}_N\left[a_{ij}\right]\right]^2+\sum_{1\le i<j\le n}\widehat{\mathbb{E}}_N\left[a_{ij}\right]\left(1-\widehat{\mathbb{E}}_N\left[a_{ij}\right]\right)\\ &-\sum_{1\le i<j\le n}\sum_{\substack{1\le i'<j'\le n\\(i,j)\neq(i',j')}}\widehat{\mathbb{E}}_N\left[a_{ij}\right]\widehat{\mathbb{E}}_N\left[a_{i'j'}\right]-\widehat{\mathbb{E}}_N\left[\rho_{ij,i'j'}\right]\\ &+4\sum_{(i,j)\in\mathcal{E}(\boldsymbol{B})}\sum_{(i',j')\in\overline{\mathcal{E}}(\boldsymbol{B})}\widehat{\mathbb{E}}_N\left[a_{ij}\right]\widehat{\mathbb{E}}_N\left[a_{i'j'}\right]-\widehat{\mathbb{E}}_N\left[\rho_{ij,i'j'}\right]\end{aligned}\tag{63}$$

In the following, we consider all the terms in (63), one at a time.

Using (62), and for N large enough, the first term in (63) is given by

$$\left[\sum_{(i,j)\in\mathcal{E}(\boldsymbol{B})}\left(1-2\widehat{\mathbb{E}}_N\left[a_{ij}\right]\right)+\sum_{1\le i<j\le n}\widehat{\mathbb{E}}_N\left[a_{ij}\right]\right]^2=\left[\sum_{(i,j)\in\mathcal{E}(\boldsymbol{B})}\left(1-2p_{ij}\right)+\sum_{1\le i<j\le n}p_{ij}\right]^2+O\left(\frac{1}{\sqrt{N}}\right).\tag{64}$$

with high probability. Also, we have the following, with high probability,

$$\sum_{1\le i<j\le n}\widehat{\mathbb{E}}_N\left[a_{ij}\right]\left(1-\widehat{\mathbb{E}}_N\left[a_{ij}\right]\right)=\sum_{1\le i<j\le n}p_{ij}\left(1-p_{ij}\right)+O\left(\frac{1}{\sqrt{N}}\right).\tag{65}$$

The last two terms in (63) can be neglected since,

$$\begin{aligned}\sum_{1\le i<j\le n}\sum_{\substack{1\le i'<j'\le n\\(i,j)\neq(i',j')}}\left[\widehat{\mathbb{E}}_N\left[a_{ij}\right]\widehat{\mathbb{E}}_N\left[a_{i'j'}\right]-\widehat{\mathbb{E}}_N\left[\rho_{ij,i'j'}\right]\right]&=\sum_{1\le i<j\le n}\sum_{\substack{1\le i'<j'\le n\\(i,j)\neq(i',j')}}\left[p_{ij}p_{i'j'}-p_{ij}p_{i'j'}\right]+O\left(\frac{1}{\sqrt{N}}\right)\\ &=O\left(\frac{1}{\sqrt{N}}\right),\end{aligned}\tag{66}$$

with high probability. Similarly

$$\sum_{(i,j)\in\mathcal{E}(\boldsymbol{B})}\sum_{(i',j')\in\overline{\mathcal{E}}(\boldsymbol{B})}\widehat{\mathbb{E}}_N\left[a_{ij}\right]\widehat{\mathbb{E}}_N\left[a_{i'j'}\right]-\widehat{\mathbb{E}}_N\left[\rho_{ij,i'j'}\right]=O\left(\frac{1}{\sqrt{N}}\right).\tag{67}$$

with high probability. Substituting (64), (65), (66), and (67) into (63) yields the following estimate for $\widehat{F}_2(\boldsymbol{B})$,

$$\widehat{F}_2(\boldsymbol{B})=\left[\sum_{(i,j)\in\mathcal{E}(\boldsymbol{B})}\left(1-2p_{ij}\right)+\sum_{1\le i<j\le n}p_{ij}\right]^2+\sum_{1\le i<j\le n}p_{ij}\left(1-p_{ij}\right)+O\left(\frac{1}{\sqrt{N}}\right)\tag{68}$$

which holds with high probability. For large N, we can neglect the $O\left(1/\sqrt{N}\right)$ term and minimize the main term, which matches the Fréchet function for the population mean. □