Properties of Latent Variable Network Models

Riccardo Rastelli[1,2,*], Nial Friel[1,2], and Adrian E. Raftery[3]

[*] riccardo.rastelli@ucdconnect.ie
[1]School of Mathematical Sciences, University College Dublin, Ireland;
[2]Insight: Centre for Data Analytics, University College Dublin, Ireland;
[3]Department of Statistics, University of Washington, Seattle, USA.

June 26, 2015

Abstract

We derive properties of Latent Variable Models for networks, a broad class of models that includes the widely-used Latent Position Models. These include the average degree distribution, clustering coefficient, average path length and degree correlations. We introduce the Gaussian Latent Position Model, and derive analytic expressions and asymptotic approximations for its network properties. We pay particular attention to one special case, the Gaussian Latent Position Models with Random Effects, and show that it can represent the heavy-tailed degree distributions, positive asymptotic clustering coefficients and small-world behaviours that are often observed in social networks. Several real and simulated examples illustrate the ability of the models to capture important features of observed networks.

Keywords: Fitness models, Latent Position Models, Latent Variable Models, Social networks, Random graphs.

1 Introduction

Networks are tools for representing relations between entities. Examples include social networks, such as acquaintance networks (Amaral et al. 2000), collaboration networks (Newman 2001) and interaction networks (Perry and Wolfe 2013), technological networks such as the World Wide Web (Albert et al. 1999), and biological networks such as neural networks (Watts and Strogatz 1998), food webs (Williams and Martinez 2000), and protein-protein interaction networks (Raftery et al. 2012).

Social networks, specifically, tend to exhibit transitivity (Newman 2003a), clustering, homophily (Newman and Park 2003), the scale-free property (Newman 2002b) and small-world behaviours (Watts and Strogatz 1998).

Networks are typically modelled in terms of random graphs. The set of nodes is fixed, and a probability distribution is defined over the space of all possible sets of edges, thereby considering the observed network as a realisation of a random variable.

One way to study networks is to define a simple generative mechanism that captures some important basic properties, such as the degree distribution (Newman et al. 2001), clustering (Newman 2009), or small-world behaviour (Watts and Strogatz 1998). These models are deliberately made simple so to be easily fitted and studied. Theoretical tractability can allow the asymptotic properties of the fitted models to be assessed, and this can give help to determine how well the models might fit real large networks. It can also allow the relationships between statistics measuring clustering, power-law behaviour and small-world behaviour to be assessed (Kiss and Green 2008; Newman 2009; Watts and Strogatz 1998).

On the other hand, various statistical models have been proposed, including Exponential Random Graph Models (Frank and Strauss 1986; Caimo and Friel 2011; Krivitsky and Handcock 2014), Latent Stochastic Blockmodels (Nowicki and Snijders 2001; Latouche et al. 2011; Airoldi et al. 2008), and Latent Position Models (Hoff et al. 2002; Raftery et al. 2012). These try to capture all the main features of observed networks within a unified framework. However, due to their more complicated structure, only limited research has been carried out to assess their properties (Daudin et al. 2008; Channarond et al. 2012; Ambroise and Matias 2012; Mariadassou and Matias 2015). Moreover, recent developments (Chatterjee and Diaconis 2013; Shalizi and Rinaldo 2013; Schweinberger and Handcock 2015) have shed light on some important limitations of ERGMs, questioning their suitability as statistical models for networks.

In this paper, we attempt to fill this gap by deriving theoretical properties of a wide family of network models, which we call Latent Variable Models (LVMs). This family includes one well-known class of statistical network models as a special case, namely the Latent Position Models (LPM) (Hoff et al. 2002; Handcock et al. 2007; Krivitsky et al. 2009). These are defined by associating an observed latent position in Euclidean space with each node, and postulating that nodes that are closer are more likely to be linked, with the probability of connection depending on the distance, typically through a logistic regression model. In the last decade, LPMs and their extensions have been widely used for applications such as the analysis of international investment (Cao and Ward 2014), trophic food webs (Chiu and Westveld 2011, 2014), signal processing (Wang et al. 2014), and education research (Sweet et al. 2013).

Analytic expressions for the clustering properties of this model in its original form are hard to derive. Because of this, we propose a new but closely related model, the Gaussian Latent Position Model. This yields simple analytic expressions or asymptotic approxima-

tions for several important clustering properties, including a complete characterisation of the degree distribution, the clustering coefficient, and the distribution of path lengths. The availability of analytic expressions facilitates the analysis of very large graphs since, for example, simulation is not required.

One result is that the Gaussian LPM can represent transitivity asymptotically, because its clustering coefficient can be asymptotically non-zero, unlike the Erdős-Rényi and Exponential Random Graph Models, whose clustering coefficient converges to zero.

One implication of our results is that the Latent Position Model in its original form cannot represent heavy-tailed degree distributions, such as power-law behaviour, or small-world behaviour, as measured by the average path length. As a result, we introduce the Gaussian Latent Position Model with Random Effects (LPMRE), and show that it can overcome these limitations and capture important features of large-size real networks. These results suggest that the Gaussian LPMRE may be a good model for social networks.

The paper is organised as follows. In Section 2 the notation is set and the main models of interest are defined. Section 3 gives the core theoretical results used in the paper. Section 4 makes use of such results to further analyse important features of LPMs, such as transitivity, homophily, scale-free properties and small-world behaviours. In Section 5, the appealing properties of Gaussian LPMREs are illustrated through empirical studies and examples. Section 6 provides several real data studies, while Section 7 concludes the paper with some final remarks.

2 Latent Variable Network Models

2.1 Notation and model assumptions

Here we introduce our notation and define the various latent variable models for networks that we consider.

A1. $\mathcal{G} = (V, E)$ is a binary random graph where V is the set of node labels and E is the set of random edges. The observed data consist of a realisation of $\mathcal{G}$. We denote $V = \{1, \dots, n\}$ and represent the observed edges through the adjacency matrix $\mathbf{Y} = \{y_{ij}\}_{(i,j)\in V\times V}$, where:

$$y_{ij} = \begin{cases} 1, & \text{if an edge from } i \text{ to } j \text{ appears in the graph,} \\ 0, & \text{otherwise.} \end{cases} \tag{2.1}$$

Furthermore we assume that edges are undirected and self-edges are not allowed, i.e. $y_{ij} = y_{ji},\ \forall (i,j) \in \tilde{V} := \{(i,j) : 1 \le i < j \le n\}$ and $y_{ii} = 0,\ \forall i \in V$, respectively. Our

analysis can easily be extended to the case of directed edges, however.

A Latent Variable Model (LVM) for networks is defined by associating an unobserved random variable $\mathbf{Z}_i \in \mathcal{Z}$ to actor i, $\forall i \in V$, for some discrete or continuous set $\mathcal{Z}$. The set of quantities $P = \{z_1, \ldots, z_n\}$ denotes a realisation of the corresponding random process.

A2. The latent variables $\mathbf{Z}_1, \ldots, \mathbf{Z}_n$ are independent and identically distributed, where each $\mathbf{Z}$ is distributed according to the probability measure $p(\,\cdot\,)$.

A3. Edges are assumed to be conditionally independent given the latent variables. Thus $\forall (i,j) \in \tilde{V}$, Y_{ij} is a Bernoulli random variable such that

$$Pr\left(Y_{ij} = 1 | \mathbf{z}_i, \mathbf{z}_j\right) = 1 - Pr\left(Y_{ij} = 0 | \mathbf{z}_i, \mathbf{z}_j\right) = r\left(\mathbf{z}_i, \mathbf{z}_j\right). \tag{2.2}$$

The modelling assumptions **A1-A3** are very general, and in fact various models of interest satisfy these, including the Random Connection Models of Meester (1996), the Fitness models of Caldarelli et al. (2002); Söderberg (2002), the LPMs of Hoff et al. (2002); Handcock et al. (2007); Krivitsky et al. (2009), and the Stochastic Blockmodel of Nowicki and Snijders (2001), among others. We now give more specific modelling assumptions that characterise Latent Position Models.

A4. In the LPM, the realised latent variables $\mathbf{Z}_i$ in **A2** are points in the Euclidean space $\mathbb{R}^d$, for a fixed d, and they are normally distributed:

$$p\left(P|\gamma\right) = \prod_{i=1}^{n} f_d\left(\mathbf{z}_i; \mathbf{0}, \gamma\right) = \prod_{i=1}^{n} (2\pi\gamma)^{-\frac{d}{2}} \exp\left\{-\frac{1}{2\gamma}\mathbf{z}_i^t \mathbf{z}_i\right\}. \tag{2.3}$$

In (2.3), γ is a positive real parameter and $f_d(\,\cdot\,; \boldsymbol{\mu}, \gamma)$ is the multivariate Gaussian density function with parameters $\boldsymbol{\mu}$ (mean) and $\gamma \mathbb{I}_d$ (covariance), where $\mathbb{I}_d$ is the $d \times d$ identity matrix and $\mathbf{A}^t$ denotes the transpose of the matrix or vector $\mathbf{A}$.

A5. In our specification of the LPM, the Gaussian LPM, the Bernoulli parameters in **A3** are given by:

$$r\left(\mathbf{z}_i, \mathbf{z}_j\right) = \tau \exp\left\{-\frac{\left(\mathbf{z}_i - \mathbf{z}_j\right)^t \left(\mathbf{z}_i - \mathbf{z}_j\right)}{2\varphi}\right\}, \tag{2.4}$$

where $\varphi > 0$, $\tau \in [0, 1]$.

Assumption **A5** is slightly different from the original formulation of the LPM of Hoff et al. (2002), in that the logistic connection function for the edges has been replaced by a non-normalised Gaussian density. The reasoning behind this choice will be addressed in Section 2.2.

A6. In the Logistic LPM of Hoff et al. (2002), the Bernoulli parameters in **A3** are given by:

$$r\left(\mathbf{z}_i, \mathbf{z}_j\right) = \frac{\exp\left\{\alpha - \beta d\left(\mathbf{z}_i, \mathbf{z}_j\right)\right\}}{1 + \exp\left\{\alpha - \beta d\left(\mathbf{z}_i, \mathbf{z}_j\right)\right\}}, \tag{2.5}$$

where $\alpha \in \mathbb{R},\ \beta > 0$ and $d\left(\mathbf{z}_i, \mathbf{z}_j\right)$ is the Euclidean distance between the latent positions $\mathbf{z}_i$ and $\mathbf{z}_j$.

2.1.1 Extensions of Latent Position Models

Two major extensions of the LPMs of Hoff et al. (2002) are Handcock et al. (2007) and Krivitsky et al. (2009). In the former, clustering is introduced through a mixture distribution on the latent process for nodal positions, while in the latter, nodal random effects are introduced to capture degree heterogeneity. In a similar fashion we introduce two variations of **A4** and **A5** to characterise the two cases:

A7. The latent positions are distributed according to a finite mixture of Gaussian distributions, i.e.:

$$p\left(P|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\gamma}, G\right) = \prod_{i=1}^{n}\left[\sum_{g=1}^{G} \pi_g f_d\left(\mathbf{z}_i; \boldsymbol{\mu}_i, \gamma_i\right)\right] \tag{2.6}$$

where $\boldsymbol{\pi}$ are the mixture weights, $\boldsymbol{\mu}$ and $\boldsymbol{\gamma}$ are the parameters for the components and G is the number of groups. The components are all assumed to arise from densities with circular contours, but possibly different volumes.

A8. For every node $s \in V$, the latent information $\tilde{\mathbf{z}}_s$ is composed of the realisation of a random latent position $\mathbf{Z}_s$, which is distributed according to $p(\,\cdot\,)$, and a random effect φ_s. This random effect is independent of $\mathbf{Z}_s$ and is distributed according to an Inverse Gamma distribution with parameters β_0 and β_1. Also, the connection probability is modified as follows:

$$Pr\left(Y_{ij} = 1|\mathbf{z}_i, \mathbf{z}_j, \varphi_i, \varphi_j, \tau\right) = \tau \exp\left\{-\frac{1}{2\left(\varphi_i + \varphi_j\right)^2}(\mathbf{z}_i - \mathbf{z}_j)^t(\mathbf{z}_i - \mathbf{z}_j)\right\}. \tag{2.7}$$

We call this the Gaussian Latent Position Model with Random Effects, or Gaussian LPMRE.

Different combinations of assumptions **A1**-**A8** generate different Latent Variable Models. The main cases considered in the present paper are summarised in Table 1.

Table 1: Latent Variable Models considered in the present paper. Latent variables are omitted from model parameters.

Notation	Description	Assumptions	Model parameters
LVM	Latent Variable Model	**A1**-**A3**	unspecified
Logistic LPM	LPM of Hoff et al. (2002)	**A1**-**A4**, **A6**	α, β, γ
Gaussian LPM	Gaussian connection LPM	**A1**-**A5**	τ, φ, γ
Gaussian LPCM	Clustering LPM	**A1**-**A3**, **A5**, **A7**	$\tau, \varphi, \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\gamma}, G$
Gaussian LPMRE	1-cluster with random effects	**A1**-**A4**, **A8**	$\tau, \boldsymbol{\varphi}, \gamma, \beta_0, \beta_1$

2.2 Motivation for the Gaussian likelihood assumption

The Logistic LPM has been widely used in network models. Assumption **A5** introduces a new function to define the probability of edges, which is proportional to a non-normalised Gaussian density. Other variations in the form of the likelihood function have been proposed in the statistical community (Gollini and Murphy 2014), but the reasoning behind the Gaussian function mainly comes from the physics literature (Deprez and Wüthrich 2013; Penrose 1991; Meester 1996). The main advantage of using the Gaussian function in place of the Logistic function is that it makes it easier to derive theoretical properties without much changing the generative process of the networks.

In the Gaussian function the model parameters τ and φ appear. The role of τ is to control the sparsity in the network, and to allow for the fact that nodes having the same latent position might not be connected.

The parameter φ encompasses the core idea of the LPM, relating the probability of edges to the distance between latent positions. Indeed, the larger the parameter φ the more long range edges are supported. Moreover, as φ goes to infinity, the model degenerates to an Erdős-Rényi random graph with connection probability τ.

Essentially, the difference between the two assumptions reduces to the fact that, as a function of the distance between nodes, the slopes of the curves are different (Figure 2.1). Even though an equivalence result is not provable, we argue that the properties of the Gaussian LPM are comparable and analogous to those of the Logistic LPM.

3 Theoretical results

In this section, we provide several theoretical results about LVMs, describing the distributions of features of networks realised from such models.

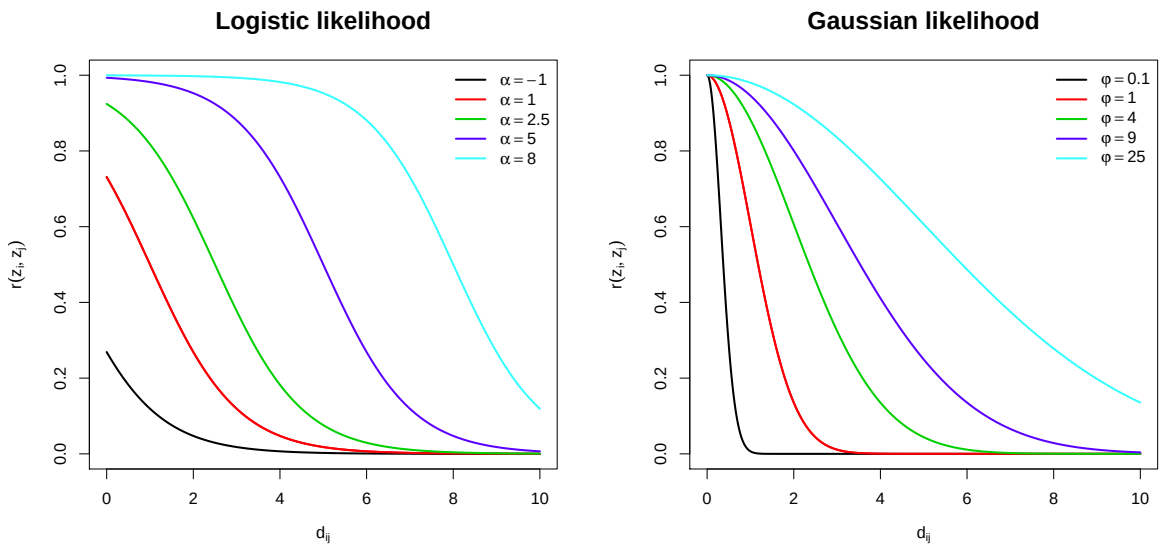

Figure 2.1: Comparison between the Logistic and Gaussian connection functions, with $\tau = \gamma = 1$. As a function of the distance between the nodes, the likelihood of a connection in both cases reaches its maximum when the distance is null, and decreases to zero as the distance increases.

3.1 Properties of the degrees

The degree of an arbitrary actor s is a discrete random variable defined by $D_s = \sum_{j \in V} Y_{sj}$. In this subsection, the properties of the degrees are characterised, describing their mixing behaviour and the distribution of the degree of a randomly chosen node, identified by the vector $\mathbf{p} = (p_0, \ldots, p_{n-1})$, where $p_k = Pr\left(D = k\right),\ \forall k = 0, \ldots, n-1$. To study the degree distribution of general LVMs (including LPMs), we propose a framework resembling that of Newman et al. (2001), which relies on the use of Probability Generating Functions (PGFs).

The study will focus on the following quantities:

- **D1:** $\theta\left(\mathbf{z}_s\right)$, defined as the probability that an actor chosen at random is a neighbour of a node with latent information $\mathbf{z}_s$.
- **D2:** The PGF of the degree of a randomly chosen actor, $G(x) = \sum_{k=0}^{n-1} x^k p_k$.
- **D3:** The factorial moments of the degree of a randomly chosen actor. Note that central and non-central moments can be recovered iteratively from factorial moments.
- **D4:** The first factorial moment, i.e. the average degree of a random node: $\bar{k}$.
- **D5:** The values of p_k, for every $k = 0, \ldots, n-1$.
- **D6:** $\bar{k}(\mathbf{z}_s)$, defined as the average degree of a node with latent information $\mathbf{z}_s$.

- **D7:** $\bar{k}_{nn}(\mathbf{z}_s)$, defined as the average degree of the neighbours of a node with latent information $\mathbf{z}_s$.
- **D8:** $\bar{k}_{nn}(k)$, defined as the average degree of the neighbours of a node with degree k.

The following main result characterises all of the quantities listed under a very general LVM:

Theorem 1. *Under assumptions **A1** − **A3**, the following results hold:*

$$\textbf{\textit{D1:}}\ \theta(\boldsymbol{z}_s) = \int_{\mathcal{Z}} p\left(\boldsymbol{z}_j\right) r\left(\boldsymbol{z}_s, \boldsymbol{z}_j\right) d\boldsymbol{z}_j \tag{3.1}$$

$$\textbf{\textit{D2:}}\ G(x) = \int_{\mathcal{Z}} p\left(\boldsymbol{z}_s\right) \left[x\theta(\boldsymbol{z}_s) + 1 - \theta(\boldsymbol{z}_s)\right]^{n-1} d\boldsymbol{z}_s \tag{3.2}$$

$$\textbf{\textit{D3:}}\ \frac{\partial^r G}{\partial x^r}(1) = \frac{(n-1)!}{(n-r-1)!} \int_{\mathcal{Z}} p\left(\boldsymbol{z}_s\right) \theta(\boldsymbol{z}_s)^r d\boldsymbol{z}_s \tag{3.3}$$

$$\textbf{\textit{D4:}}\ \bar{k} = (n-1) \int_{\mathcal{Z}} p\left(\boldsymbol{z}_s\right) \theta(\boldsymbol{z}_s) d\boldsymbol{z}_s \tag{3.4}$$

$$\textbf{\textit{D5:}}\ p_k = \int_{\mathcal{Z}} p\left(\boldsymbol{z}_s\right) \binom{n-1}{k} \theta(\boldsymbol{z}_s)^k \left[1 - \theta(\boldsymbol{z}_s)\right]^{n-k-1} d\boldsymbol{z}_s \tag{3.5}$$

$$\textbf{\textit{D6:}}\ \bar{k}(\boldsymbol{z}_s) = (n-1)\theta(\boldsymbol{z}_s) \tag{3.6}$$

$$\textbf{\textit{D7:}}\ \bar{k}_{nn}(\boldsymbol{z}_s) = 1 + \frac{(n-2)}{\theta\left(\boldsymbol{z}_s\right)} \int_{\mathcal{Z}} p\left(\boldsymbol{z}_j\right) r\left(\boldsymbol{z}_s, \boldsymbol{z}_j\right) \theta\left(\boldsymbol{z}_j\right) d\boldsymbol{z}_j \tag{3.7}$$

$$\textbf{\textit{D8:}}\ \bar{k}_{nn}(k) = \frac{1}{p_k} \int_{\mathcal{Z}} p(\boldsymbol{z}_j) \binom{n-1}{k} \theta(\boldsymbol{z}_j)^k \left[1 - \theta(\boldsymbol{z}_j)\right]^{n-k-1} \bar{k}_{nn}(\boldsymbol{z}_j) d\boldsymbol{z}_j \tag{3.8}$$

The proof of Theorem 1 is provided in Appendix A.1.

Remark. Equation (3.8) is a generalisation of a result from Boguná and Pastor-Satorras (2003), where a general framework to study the degree correlations for the fitness model of Caldarelli et al. (2002) and Söderberg (2002) is introduced.

Remark. Particular instances of some of the results of Theorem 1 have been already shown in Olhede and Wolfe (2012) and Channarond et al. (2012); Daudin et al. (2008) for Stochastic Block Models and Fitness models, without resorting to PGFs. Theorem 1 encompasses those special cases and extends the range of results offered.

The results presented in Theorem 1 are valid for all LVMs. Essentially, they relate the distributional assumptions about the latent variables and the edge probabilities to the properties of the degrees of the realised networks.

We now apply these results to LPMs. The following Corollaries show how the formulas involved in **D1**-**D8** simplify under the Gaussian models of Table 1. Proofs are shown in Appendices A.1.1 and A.1.2.

Corollary 1. *Under the Gaussian LPM, the following quantities have an explicit form:*

$$\textbf{\textit{D1:}}\ \theta(\boldsymbol{z}_s) = \tau \left(\frac{\varphi}{\gamma+\varphi}\right)^{\frac{d}{2}} \exp\left\{-\frac{1}{2(\gamma+\varphi)}\boldsymbol{z}_s^t \boldsymbol{z}_s\right\} \tag{3.9}$$

$$\textbf{\textit{D3:}}\ \frac{\partial^r G}{\partial x^r}(1) = \frac{(n-1)!}{(n-r-1)!}\tau^r \left\{\frac{\varphi^r}{(\gamma+\varphi)^{r-1}\left[(r+1)\gamma+\varphi\right]}\right\}^{\frac{d}{2}} \tag{3.10}$$

$$\textbf{\textit{D4:}}\ \bar{k} = (n-1)\tau \left\{\frac{\varphi}{2\gamma+\varphi}\right\}^{\frac{d}{2}} \tag{3.11}$$

$$\textbf{\textit{D7:}}\ \bar{k}_{nn}(\boldsymbol{z}_s) = 1 + \bar{k}\left(\frac{n-2}{n-1}\right)\frac{f_d\left(\boldsymbol{z}_s; \boldsymbol{0}, \frac{\gamma^2+3\gamma\varphi+\varphi^2}{2\gamma+\varphi}\right)}{f_d\left(\boldsymbol{z}_s; \boldsymbol{0}, \gamma+\varphi\right)} \tag{3.12}$$

Note that $\theta(\,\cdot\,)$ has an explicit expression, thus evaluation of the quantities in **D2**, **D5** and **D8** boils down to an approximation of a single integral.

Corollary 2. *Under the Gaussian LPCM, the following results hold:*

$$\textbf{\textit{D1:}}\ \theta(\boldsymbol{z}_s) = \tau\,(2\pi\varphi)^{\frac{d}{2}} \sum_{g=1}^{G} \pi_g f_d\left(\boldsymbol{z}_s; \boldsymbol{\mu}_g, \gamma_g+\varphi\right) \tag{3.13}$$

$$\textbf{\textit{D4:}}\ \bar{k} = (n-1)\tau\,(2\pi\varphi)^{\frac{d}{2}} \sum_{g=1}^{G}\sum_{h=1}^{G} \pi_g\pi_h f_d\left(\mu_g-\mu_h; \boldsymbol{0}, \gamma_g+\gamma_h+\varphi\right). \tag{3.14}$$

Also, the degree distribution is a continuous mixture of binomial distributions, where the mixture weights are themselves distributed as mixtures of Gaussians:

$$\textbf{\textit{D7:}}\ p_k = \int_{\mathbb{R}^d} \left[\sum_{g=1}^{G} \pi_g f_d\left(\boldsymbol{z}_s; \mu_g, \gamma_g\right)\right] \binom{n-1}{k} \theta(\boldsymbol{z}_s)^k \left[1-\theta(\boldsymbol{z}_s)\right]^{n-k-1} d\boldsymbol{z}_s. \tag{3.15}$$

Under the Gaussian LPMRE, none of the equations can be written explicitly, since the integrals over the random effects cannot be evaluated analytically. However, we will make use of the following two quantities, which will be calculated in an approximate form:

$$\theta(\mathbf{z}_s, \varphi_s) = \int_{\mathbb{R}^d}\int_0^\infty f_d\left(\mathbf{z}_j; \mathbf{0}, \gamma\right) p\left(\varphi_j|\beta_0, \beta_1\right) r\left(\mathbf{z}_s, \mathbf{z}_j\right) d\varphi_j d\mathbf{z}_j, \tag{3.16}$$

$$G^{(r)}(1) = \frac{(n-1)!}{(n-r-1)!}\int_{\mathbb{R}^d}\int_0^\infty f_d\left(\mathbf{z}_s; \mathbf{0}, \gamma\right) p\left(\varphi_j|\beta_0, \beta_1\right) \theta(\mathbf{z}_s, \varphi_s)^r d\mathbf{z}_s d\varphi_s, \tag{3.17}$$

Remark. The advantage of using the Gaussian function rather than the Logistic function of Hoff et al. (2002); Handcock et al. (2007); Krivitsky et al. (2009) is mainly highlighted in Corollary 1: under the Gaussian hypothesis most of the integrals of Equations 3.1-3.8 can be evaluated analytically since they become a convolution of two Gaussian densities, which is solvable for any d. Also, quantities that do not have an exact expression, such as p_k or $\bar{k}_{nn}(k)$, can be efficiently evaluated through numerical methods, since the number of integrals to approximate is constant (depending on d, but not on n).

Remark. In Gaussian LPMs, a nonidentifiability issue arises between the parameters φ and γ, since the factorial moments depend only on their ratio, φ/γ. We argue, however, that both parameters should be included in our study, to keep the model as close as possible to the original LPM of Hoff et al. (2002), and to provide a proper basis for possible extensions, such as the Gaussian LPCM and the Gaussian LPMRE.

3.2 Clustering coefficient

In this section, we take advantage of the Gaussian assumption to study the clustering coefficient value for Gaussian LPMs analytically.

Since there is more than one definition for the clustering coefficient, we clarify that the one used in this paper is the global clustering coefficient of Newman (2003a), equal to three times the number of triangles divided by the number of connected triples of nodes. Thanks to the exchangeability of actor labels, this quantity is an unbiased estimator of the probability that, given two consecutive edges, the extremities of such 2-steps path are connected themselves.

Proposition 1. *Under assumptions **A1**-**A3**, the global clustering coefficient $\mathcal{C}$ can be written as:*

$$\mathcal{C} = \frac{\int_{\mathcal{Z}}\int_{\mathcal{Z}}\int_{\mathcal{Z}} p(\boldsymbol{z}_i)p(\boldsymbol{z}_k)p(\boldsymbol{z}_j)r\left(\boldsymbol{z}_i,\boldsymbol{z}_k\right)r\left(\boldsymbol{z}_k,\boldsymbol{z}_j\right)r\left(\boldsymbol{z}_j,\boldsymbol{z}_i\right)d\boldsymbol{z}_i d\boldsymbol{z}_k d\boldsymbol{z}_j}{\int_{\mathcal{Z}}\int_{\mathcal{Z}}\int_{\mathcal{Z}} p(\boldsymbol{z}_i)p(\boldsymbol{z}_k)p(\boldsymbol{z}_j)r\left(\boldsymbol{z}_i,\boldsymbol{z}_k\right)r\left(\boldsymbol{z}_k,\boldsymbol{z}_j\right)d\boldsymbol{z}_i d\boldsymbol{z}_k d\boldsymbol{z}_j}. \tag{3.18}$$

Under the Gaussian LPM both the numerator and the denominator can be expressed analytically, yielding the following result:

$$\mathcal{C} = \tau\left(\frac{\gamma+\varphi}{3\gamma+\varphi}\right)^{\frac{d}{2}}. \tag{3.19}$$

A proof of Proposition 1 is provided in Appendix A.4. We note that the (3.19) gives an exact result for the clustering coefficient of an LPM of any size. This is an interesting result and contrasts with many network models, where the clustering coefficient can only be recovered asymptotically. Some interesting consequences of (3.19) will be illustrated in Section 4.3.

3.3 Connectivity properties

The study of the theoretical properties of LPMs can be further extended, characterising the connectivity structure of realised networks. To do so, we give the definition of a path for a random graph, and show a general result about the connection of two nodes in Gaussian LPMs, once their latent position is known.

Definition 3.3.1 (Path)**.** Under assumptions **A1**-**A3**, a k-step path is a sequence of $k+1$ distinct nodes $\{i_0, i_1, \ldots, i_k\}$ such that an edge is present between every two consecutive nodes, i.e. $y_{i_0 i_1} = y_{i_1 i_2} = \cdots = y_{i_{k-1} i_k} = 1$.

Under the same assumptions, the probability of a k-step path appearing between two nodes with latent information $\mathbf{z}_i$ and $\mathbf{z}_j$ can be written as:

$$I_k(\mathbf{z}_i, \mathbf{z}_j) = \int_{\mathcal{Z}} \cdots \int_{\mathcal{Z}} p(\mathbf{z}_1) \ldots p(\mathbf{z}_{k-1}) r(\mathbf{z}_i, \mathbf{z}_1)\, r(\mathbf{z}_1, \mathbf{z}_2) \cdots r(\mathbf{z}_{k-1}, \mathbf{z}_j)\, d\mathbf{z}_1 \cdots d\mathbf{z}_{k-1}. \tag{3.20}$$

For a Gaussian LPM, the integrals on the right-hand side of (3.20) involve Gaussian kernels only, and therefore they can be evaluated exactly. We provide a more explicit formula for $I_k(\mathbf{z}_i, \mathbf{z}_j)$ in the following Proposition:

Proposition 2. *Under the Gaussian LPM, let $I_k(\boldsymbol{z}_i, \boldsymbol{z}_j)$ be defined as in* (3.20)*, for any $k = 1, 2, \ldots, n-1$, $\boldsymbol{z}_i \in \mathbb{R}^d$ and $\boldsymbol{z}_j \in \mathbb{R}^d$. Define the following recurrence relations:*

$$\begin{cases} h_{r+1} &= h_r \alpha_r^{-d} \tau (2\pi\varphi)^{\frac{d}{2}} f_d\left(\boldsymbol{z}_i; \boldsymbol{0}, \frac{\omega_r+\gamma}{\alpha_r^2}\right) \\ \alpha_{r+1} &= \frac{\alpha_r \gamma}{\omega_r+\gamma} \\ \omega_{r+1} &= \frac{\omega_r\varphi+\omega_r\gamma+\gamma\varphi}{\omega_r+\gamma} \end{cases}, \text{ with } \begin{cases} h_1 &= \tau (2\pi\varphi)^{\frac{d}{2}} \\ \alpha_1 &= 1 \\ \omega_1 &= \varphi \end{cases}. \tag{3.21}$$

Then, the following result holds:

$$I_k(\boldsymbol{z}_i, \boldsymbol{z}_j) = h_k f_d(\boldsymbol{z}_j - \alpha_k \boldsymbol{z}_i; \boldsymbol{0}, \omega_k), \quad \text{for } k = 1, 2, \ldots, n-1. \tag{3.22}$$

The proof of Proposition 2 is provided in Appendix A.2.

Remark. Note that the previous result could be extended by integrating out the latent positions $\mathbf{z}_i$ and $\mathbf{z}_j$ as well. However, this is not of interest for the present work.

The result of Proposition 2 is a useful tool for studying the statistical properties of path lengths for Gaussian LPMs, which we develop in Section 4.4.

4 Properties of realised networks

We now use the results in the previous section to obtain properties of the Gaussian LPM.

A main drawback of all LPMs is that, given the complete set of latent positions, the evaluation of the likelihood for the corresponding realised graph requires the calculation of a distance matrix, with a computational and storage cost of $O(n^2)$. This cost is the main obstacle to inference for large graphs, making estimation impractical for networks larger than a few thousands nodes. The issue extends also to the generation of LPMs, which is usually performed in two sequential steps: firstly latent positions are sampled, and then edges are created with the Gaussian probability. The evaluation of the distance

matrix is thus needed in between the two steps. This makes any empirical study of the properties of LPMs rather inefficient and limited to relatively small graphs, only.

By contrast, the results presented in Theorem 1 and related Corollaries involve either exact formulas, which have negligible computational cost, or integral approximations whose computational cost is independent of n. Hence, the analysis that we propose in this Section does not require any intensive calculation and can be performed on networks of any size. Note that Raftery et al. (2012) proposed a computational approximation to overcome this difficulty, whereas here we provide exact results and analytical approximations.

4.1 Characterisation of the degree distribution for LPMs

Empirical evaluations (Newman 2003b) suggest that typically the proportion of nodes with degree greater than k is expected to be proportional to $k^{-\alpha}$, for a positive α which can be as small as 2. Networks exhibiting such behaviour are usually referred to as scale-free, and the corresponding degree distribution is said to follow a power-law decay. The highly connected nodes, denoted hubs, fulfil a crucial role in defining the structure of the network (Albert et al. 2000), and as a result this is a feature which many network models aim to capture (Barabási and Albert 1999; Newman et al. 2001).

According to the results of the previous section, the theoretical degree distribution of a Gaussian LPM has the form of a continuous mixture of binomials, and can be approximated efficiently for any network size. Figure 4.1 shows approximate degree distributions for various choices of model parameters.

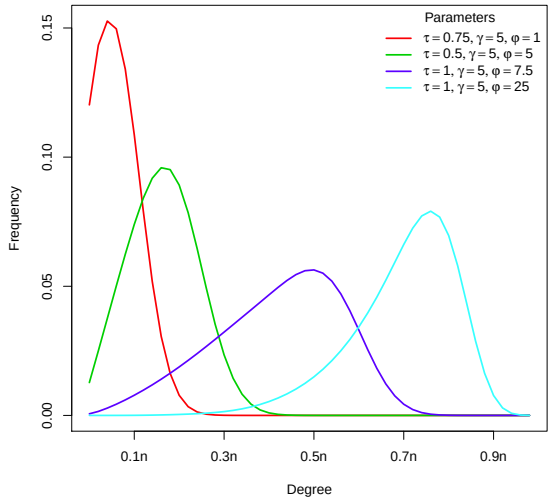

Figure 4.1: Gaussian Latent Position Model: Approximate degree distribution for different sets of model parameters τ, γ, φ.

While the degree distributions of sparse networks often resemble Poisson distributions, denser networks tend to be associated with more left-skewed shapes. However, the theoretical degree distribution of LPMs in Figure 4.1 resembles a truncated shape, suggesting that the model may not successfully represent heavy tails. It should be noted, however, that truncated shapes do arise in social networks: data are often collected through surveys, where each actor is asked to specify up to a fixed number of preferences, so that the degree distribution will exhibit an artificial truncation at the corresponding value. Popular social datasets have been obtained using such a design, such as Sampson's monks data (Sampson 1968) and the Adolescent Health data (Handcock et al. 2007). Moreover, some important empirical evidence has been shown in Dunbar (1992) demonstrating the existence of a theoretical cognitive limit on the number of stable relationships that social actors can maintain. Hence both power-laws and non-power-laws behaviours are of interest in statistical modelling of networks.

We now propose a more rigorous analysis of the degree distribution using the dispersion and skewness indexes, which can be evaluated through the exact formulas for the factorial moments in (3.10).

Corollary 3. *Under the Gaussian LPM, the dispersion index is given by:*

$$\mathcal{D} = 1 + (n-2)\tau \left(\frac{\varphi(2\gamma+\varphi)}{(\gamma+\varphi)(3\gamma+\varphi)} \right)^{\frac{d}{2}} - (n-1)\tau \left(\frac{\varphi}{2\gamma+\varphi} \right)^{\frac{d}{2}}. \tag{4.1}$$

The proof is given in Appendix A.3.

Remark. The calculation of the skewness does not involve any simplification, and so it is omitted here.

The dispersion index can be used to assess how dispersed the distribution is when compared to a Poisson, which has an index of 1. A value greater than 1 corresponds to an overdispersed distribution while a value smaller than 1 corresponds to an underdispersed one. The Binomial distribution arising from a finite Erdős-Rényi random graph has a dispersion index smaller than 1, and so it qualifies as underdispersed.

Corollary 3 allows us to study how the model parameters τ, γ and φ affect the dispersion of the distribution. For $d = 2$, our results can be summarised as follows:

- When $\varphi = \gamma(\sqrt{n-1} - 2)$, the distribution has dispersion index 1, typical of a Poisson distribution.
- When $\varphi < \gamma(\sqrt{n-1} - 2)$, the distribution has dispersion index greater than 1, so that the distribution is overdispersed.
- When $\varphi > \gamma(\sqrt{n-1} - 2)$, the distribution has dispersion index smaller than 1, typical of a Binomial distribution, and so is underdispersed.

Note that the characterisation does not depend on τ.

The left panel of Figure 4.2 shows the dispersion as a function of the model parameters. The motivation behind this result is that the Erdős-Rényi random graph model is recovered as a special case asymptotically, as φ gets larger. Therefore, as φ increases, the model degenerates and the degree distribution becomes binomial and thus underdispersed, regardless of how sparse the network is. If φ is small enough, namely below the threshold, then the model is nondegenerate and produces networks with an overdispersed degree distribution. Hence, Gaussian LPMs are able to represent degree heterogeneity, since for many choices of the model parameters the degree distribution is overdispersed. However, degree heterogeneity does not imply heavy tails or power-law behaviour.

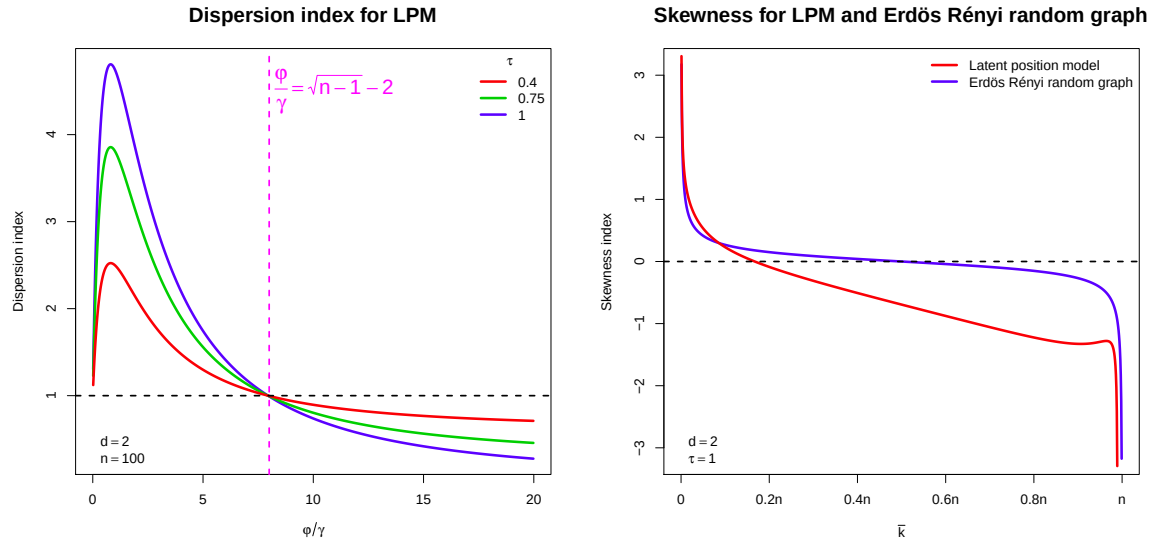

Figure 4.2: Gaussian Latent Position Model: ***Left:*** *Dispersion index versus the ratio between φ and γ. The vertical line is the threshold corresponding to a Poisson dispersion. For larger values of φ, the distributions arising are not more dispersed than an Erdős-Rényi random graph, asymptotically degenerating to such model as φ gets larger.* ***Right:*** *Unless the graph is very sparse, the skewness index for Gaussian LPMs (red line) is smaller than the skewness of a Erdős-Rényi random graph (blue line) with the same average degree.*

We now analyse the skewness index, which is useful for identifying asymmetries in overdispersed distributions. In the case of degree distributions of networks, a negative value of the skewness index corresponds to shapes exhibiting a left tail heavier than the right one, while a positive value corresponds to the opposite behaviour. As a tool to assess the presence of hubs, we expect a scale-free network to have a positive and relatively large skewness index. However, as shown in the right panel of Figure 4.2, this scenario does not arise in Gaussian LPMs.

Given that in Erdős-Rényi random graphs p_k goes to zero at the rate $1/k!$ (i.e. power laws are not represented), the right panel in Figure 4.2 shows that, unless the graph is

very sparse, Gaussian LPMs exhibit degree distributions that are always more skewed to the left than those of the Erdős-Rényi model with the same average degree. Even for very sparse networks, the difference is not large enough to justify the presence of a low-order power-law tail.

This shows that Gaussian LPMs cannot capture power-law behaviour. They are able to represent degree heterogeneity, but in the sense that degrees will not be concentrated around the mean value, but will rather have a nontrivially dispersed distribution between 0 and a maximum degree value, confirming the shapes already shown in Figure 4.1.

4.2 Degree correlations

In the study of networks, one is often interested in the mixing properties of the graph. One mixing structure arises when nodes that share common features are more likely to be linked. In the context of social networks, this behaviour is called homophily.

A special case is mixing according to the nodes' degrees, called degree correlation. For example, one might be interested in whether the degrees of two random nearest neighbours are positively or negatively correlated. Positive correlation, or assortative mixing of the degrees, is a recurring feature in social networks (Newman and Park 2003; Newman 2002a), in contrast to many other kinds of networks (World Wide Web, protein interactions, food webs; see Newman (2003b)), which typically have negative degree correlation or dissortative mixing.

Here, we illustrate the fact that Gaussian LPMs can represent assortative mixing in the degrees, using the results of Theorem 1. Equation (3.12) shows that the Average Nearest Neighbours' Degree (ANND) of an arbitrary node i is an exact function of its latent position $\mathbf{z}_i$. The left panel of Figure 4.3 displays this function in terms of the distance between $\mathbf{z}_i$ and the centre of the latent space.

It is not surprising that nodes located closer to the centre have highly connected neighbours. Instead, (3.8) provides a less explicit formula for the ANND index as a function of the degree of node i, rather than its distance from the centre. This quantity can be efficiently approximated for every degree value. The right panel of Figure 4.3 represents this case. The average degree of the neighbours of a node of degree k, $\bar{k}_{nn}(k)$, appears to be a nondecreasing function of the degree k, indicating the presence of assortative mixing in the degrees, using the same criterion as Boguná and Pastor-Satorras (2003). It follows that realised Gaussian LPM networks exhibit assortative mixing of the degrees, suggesting them to be well suited for social networks (Newman and Park 2003).

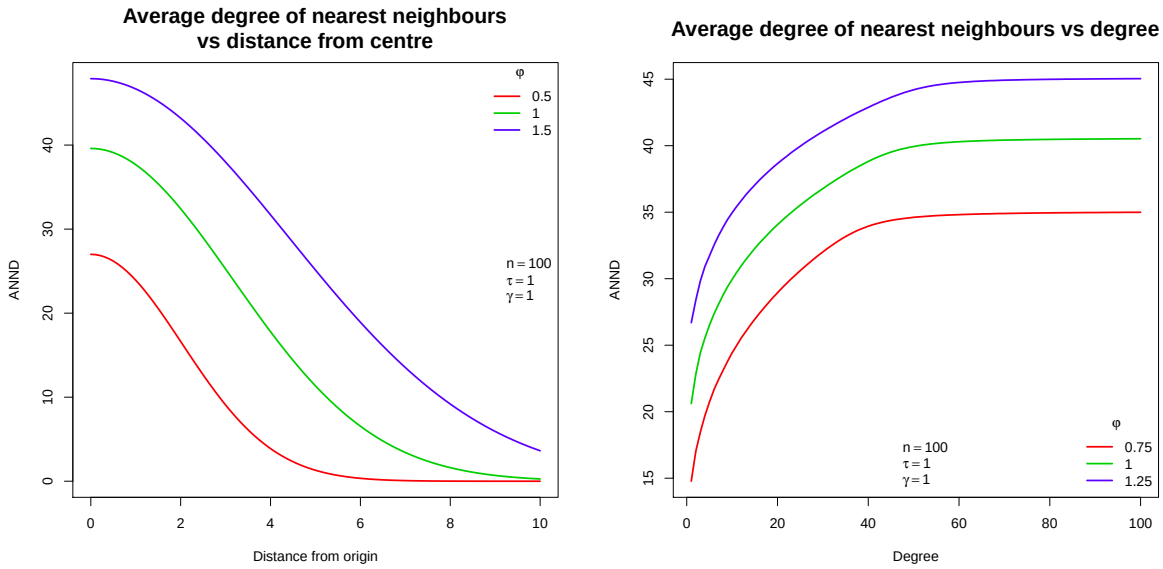

Figure 4.3: Gaussian Latent Position Model: ***Left:*** *Average degree of the closest neighbours of a node as a function of its distance from the centre. Nodes located in the centre will more likely connect to high degree nodes.* ***Right:*** *Average degree of the closest neighbours as a function of the degree of a node. The ANND index is clearly a nondecreasing function, verifying that Gaussian LPMs exhibit assortative mixing in the degrees of the nodes.*

4.3 Asymptotics for the clustering coefficient

Transitivity, defined as the propensity of two neighbours of a node also to be neighbours of one another, is ubiquitous in network analysis. In social networks, the tendency of three or more nodes to cluster is a feature of interest since it has a nontrivial relation with the structure of path lengths, for example impacting the dynamics of the spread of diseases (Newman 2003a, 2009; Kiss and Green 2008).

LPMs capture transitivity in a very natural way. Indeed, when two actors have a neighbour in common, it is expected that the three corresponding nodes will be close in the latent space, making triangles more likely. This reasoning extends to higher order configurations as well. In this section, we show how Proposition 1 provides a more objective justification to this intuition.

One well-known drawback of the Erdős-Rényi model is that it cannot capture transitivity when the network is large. To see this, let p be the connection probability and $\bar{k} = p(n-1)$ be the expected average degree of the corresponding realised network. We focus on the realistic case where the size of the network increases (n tends to infinity), while $\bar{k}$ remains constant with respect to n. It follows that p must tend to zero as n increases, as well as $\mathcal{C} \to 0$ since $\mathcal{C} = p$. Hence, asymptotically, the clustering coefficient for Erdős-Rényi random graphs is zero.

Even more structured models such as Exponential Random Graph Models, have been

shown to degenerate asymptotically to Erdős-Rényi random graphs, under some nonrestrictive conditions (Chatterjee and Diaconis 2013), thus losing the ability to represent a nontrivial transitivity structure.

In contrast, Gaussian LPMs can represent transitivity, even asymptotically. To see this, first, recall (3.11), which defines the average degree of a random node in a Gaussian LPM. In order to have an asymptotically constant average degree $\bar{k}_0$, the parameters φ and γ should satisfy:

$$\varphi = \frac{2\bar{k}_0^{\frac{2}{d}}\gamma}{(n-1)^{\frac{2}{d}}\tau^{\frac{2}{d}} - \bar{k}_0^{\frac{2}{d}}}. \tag{4.2}$$

In the limit of large n, the corresponding clustering coefficient satisfies:

$$\mathcal{C} = \frac{\tau}{3^{\frac{d}{2}}}. \tag{4.3}$$

Thus the limiting clustering coefficient has a non-zero value that can be as large as $3^{-\frac{d}{2}}$. This highlights an important difference between the Erdős-Rényi and Exponential Random Graph models on one hand, and LPMs on the other, in that the latter are able to represent transitivity in large networks.

Furthermore, the non-null clustering coefficient classifies Gaussian LPMs as highly clustered networks. Such models lack the loopless tree structure which simplifies the study of component sizes and path lengths. A review of the main difficulties arising when dealing with highly clustered models can be found in Newman (2002b).

4.4 Path lengths

In a well known experiment, Milgram (1967) observed that any two strangers are connected by a chain of intermediate acquaintances of length at most six. Later on, similar observations were made in Albert et al. (1999) about the connectivity of certain portions of the Internet, stating that any two web pages are at most 19 clicks away from one another. The small-world effect defines this behaviour exactly: given any two connected nodes, the shortest path from one node to the other will have an average length which is very small when compared to the size of the network n, typically comparable to $\log(n)$ or smaller (Newman 2001). The small-world property has motivated research on the connectivity of graphs, relevant to fields such as communication systems, epidemiology and optimisation.

Hence, understanding how a statistical model relates to the small-world property is important. For LPMs, not much is known about the diameter and connectivity of the realised networks. Here, we use Proposition 2 to apply a procedure similar to that of Fronczak et al. (2004), showing how the distribution of the geodesic distances can be evaluated in a Gaussian LPM. We also characterise the average path length (APL)

for Gaussian LPM networks of any size, giving appropriate insights on the asymptotic behaviour of such an index.

Fronczak et al. (2004) focused on the family of fitness models for networks, which includes Erdős-Rényi random graphs and the preferential attachment model of Barabási and Albert (1999). These models satisfy assumptions **A1-A3**, where the latent information is coded by a fitness value h_i, for every $i \in V$. Then, edge probabilities are given by:

$$r\left(h_i, h_j\right) = \frac{h_i h_j}{\beta}, \tag{4.4}$$

where β is a suitable constant. The model degenerates to an Erdős-Rényi random graph when $h_i = \bar{k}$ for every i, and $\beta = \bar{k}(n-1)$.

Here, we exploit the fact that fitness models and LPMs both originate from LVMs, generalising the work of Fronczak et al. (2004) to a wider family of models. To study the connectivity of the networks and the path lengths' distribution, we focus on the quantities $\ell_k\left(\mathbf{z}_i, \mathbf{z}_j\right)$, defined as the probability that the shortest path between two nodes located in $\mathbf{z}_i$ and $\mathbf{z}_j$ has length k. We also define $r_k\left(\mathbf{z}_i, \mathbf{z}_j\right)$ as the probability that a path of length k exists between two nodes. In both definitions, and from now on, we condition on the fact that the two nodes are connected, i.e. that there exists a finite-length path that has the two nodes as extremes. Such an assumption is natural since usually statistics of path lengths are defined only for sets of connected nodes. Note that $I_k\left(\mathbf{z}_i, \mathbf{z}_j\right)$ differs from $r_k\left(\mathbf{z}_i, \mathbf{z}_j\right)$ in that the latter is the probability that there is at least one k-step path between the two nodes. We now describe a way to evaluate $\ell_k\left(\mathbf{z}_i, \mathbf{z}_j\right)$ efficiently, as a function of the model parameters of a Gaussian LPM.

A9. The graphs considered are dense enough, such that for every $(i, j) \in \tilde{V}$, if there exists a path of length k between nodes i and j, then a path of length t exists between the same nodes for every $t = k+1, \ldots, n-1$.

Proposition 3. *Under the Gaussian LPM and assumption **A9**, let i and j be any two nodes. Then the following two statements are equivalent:*

- *The geodesic distance between i and j is less than k.*
- *There exists a k-step path between i and j.*

The proof of Proposition 3 relies heavily on **A9** and is straightforward. From Proposition 3 it follows that, for any i and j:

$$r_k\left(\mathbf{z}_i, \mathbf{z}_j\right) = \sum_{t=1}^{k} \ell_t\left(\mathbf{z}_i, \mathbf{z}_j\right). \tag{4.5}$$

Moreover, since $\ell_1(\mathbf{z}_i, \mathbf{z}_j) = r_1(\mathbf{z}_i, \mathbf{z}_j) = r(\mathbf{z}_i, \mathbf{z}_j)$, the following holds:

$$\ell_k(\mathbf{z}_i, \mathbf{z}_j) = r_k(\mathbf{z}_i, \mathbf{z}_j) - r_{k-1}(\mathbf{z}_i, \mathbf{z}_j). \tag{4.6}$$

Hence, we aim to characterise $r_k(\mathbf{z}_i, \mathbf{z}_j)$, thereby deducing the properties of $\ell_k(\mathbf{z}_i, \mathbf{z}_j)$.

Each possible path of length k from i to j can be thought of as a Bernoulli random variable, having a success if all the edges involved in the path appear, or not having a success if any of those edges fail to appear. For an Erdős-Rényi random graph with average degree $\bar{k} = (n-1)p$, the parameter of such a random variable is p^k. For Gaussian LPMs, the success probability is $I_k(\mathbf{z}_i, \mathbf{z}_j)$, which has been characterised in Proposition 2.

However, we are interested in $r_k(\mathbf{z}_i, \mathbf{z}_j)$, which is the probability of the union of all the k-steps paths from i to j. Unfortunately, these variables are not independent, since different paths will have edges in common. We circumvent this issue by pretending that all such paths are mutually independent, following the reasoning of Fronczak et al. (2004). This assumption makes sense when k is much smaller than n. In fact, for the purpose of the study of shortest path lengths, estimates of $r_k(\mathbf{z}_i, \mathbf{z}_j)$ will be needed only for small ks, since in the general case $\ell_k(\mathbf{z}_i, \mathbf{z}_j)$ will drop to zero very quickly.

Using the results of Proposition 2 and Lemma 1 of Fronczak et al. (2004), we obtain:

$$\ell_k(\mathbf{z}_i, \mathbf{z}_j) \approx \exp\left\{-n^{k-1} I_{k-1}(\mathbf{z}_i, \mathbf{z}_j)\right\} - \exp\left\{-n^k I_k(\mathbf{z}_i, \mathbf{z}_j)\right\}. \tag{4.7}$$

Equation (4.7) gives a general formula to evaluate the distribution of the geodesic distance $\ell_k(\mathbf{z}_i, \mathbf{z}_j)$ for every $k << n$ for dense Gaussian LPM networks.

In Figure 4.4 a comparison between the empirical and theoretical values obtained is shown. The first two panels of Figure 4.4 give a representation of how close the approximation of the path length distribution can be, for a dense Gaussian LPM network and a less dense one. Note that in less dense networks the assumption that $k << n$ is less likely to hold because more sparsity will imply longer shortest paths.

Also, once $\ell_k(\mathbf{z}_i, \mathbf{z}_j)$ is known for every k, a straightforward evaluation of the APL can be obtained by averaging over all possible values of k, $\mathbf{z}_i$ and $\mathbf{z}_j$. The agreement of the estimation with the results from an empirical study is shown in the right panel of Figure 4.4. As expected, the estimation is more accurate for graphs with a higher average degree. However, the results show that such an index is more tolerant when assumptions tend to be violated, possibly because the bias is limited when values are averaged.

Figure 4.5 shows that Gaussian LPMs typically have a higher APL than corresponding Erdős-Rényi random graphs. In the left panel, the APL is plotted against the average degree of the network. It appears that the sparser the network, the more marked the difference with Erdős-Rényi random graphs is. Instead, as the network gets denser,

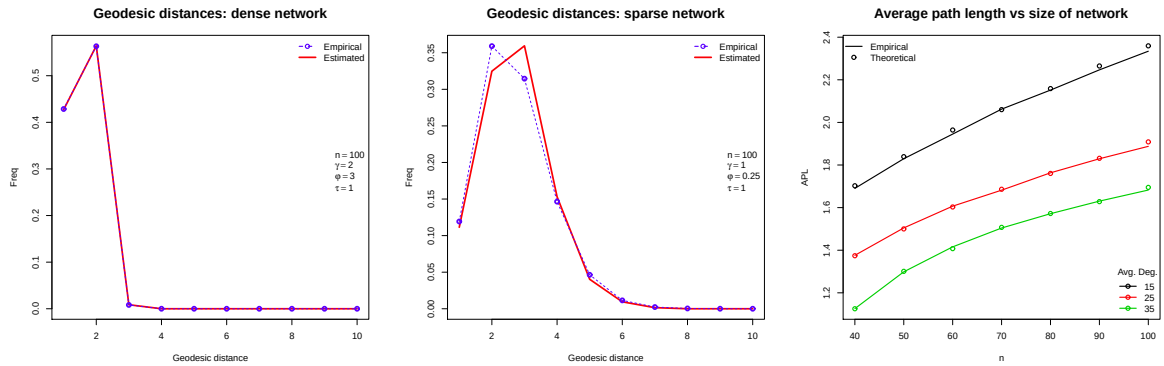

Figure 4.4: Geodesic distances and average path lengths for the Gaussian LPM model. ***Left and centre:*** *Comparison between empirical and theoretical values for the distribution of geodesic distances. Networks generated are composed of* 100 *nodes. The left panel corresponds to a more dense graph (average degree is approximately* 42*) while the one in the centre corresponds to a more sparse graph (average degree is approximately* 11*).* ***Right:*** *Comparison between empirical (lines) and theoretical (dots) values of APL. The parameters* τ *and* γ *are set to* 1*.*

Gaussian LPMs tend to behave more and more similarly to Erdős-Rényi random graphs. In the right panel of Figure 4.5, APL values are shown for larger Gaussian LPMs networks. In this case the average degree is kept constant, highlighting the asymptotic behaviour of the statistic.

APL values for the corresponding Erdős-Rényi random graphs are also shown in Figure 4.5. The Gaussian LPM networks typically have a higher APL, which grows faster than the logarithm of the size of the network.

Figure 4.6 illustrates a possible reason for this behaviour. The distance from a node to the centre of the latent space is plotted versus its geodesic distance to a second node picked at random. There is clear heterogeneity, in contrast with the behaviour of Erdős-Rényi random graphs. Clearly, when averaging over all the possible positions of the second randomly chosen node, important contributions are given by distant isolated nodes, thereby increasing the APL value.

5 Advantages of random effects models

In the previous section, we have shown that, although the Gaussian LPM can capture degree heterogeneity, it cannot represent the power-law behaviour of many observed degree distributions. In addition, the model has shortcomings in representing the small-world behaviour, in that the APL grows faster than the log of the number of actors.

In the Logistic LPM context, Krivitsky et al. (2009) addressed similar issues by adding node-specific random effects to represent different levels of social involvement. Here, we propose an extension of the Gaussian LPM (namely the Gaussian LPMRE of Table 1)

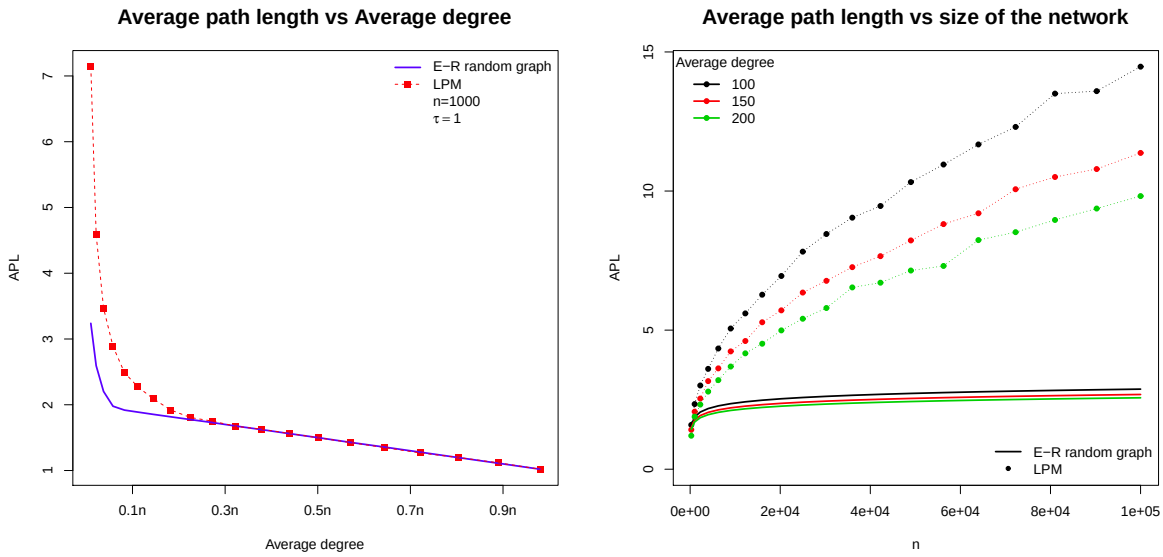

Figure 4.5: ***Left****: APL against the average degree of a* 1000 *nodes network, compared with the corresponding Erdős-Rényi random graph. The two behaviours diverge for sparse graphs, in which case Gaussian LPMs exhibit a larger APL.* ***Right****: Asymptotic behaviour for the APL is shown. Average degree of the network is kept constant while the size n is on the horizontal axis. The continuous lines represent the APL value for corresponding Erdős-Rényi random graphs with same average degrees. APL is typically higher in LPM, and grows proportionally to a function which dominates the logarithm.*

following the same reasoning.

In the Gaussian LPMRE, the connectivity parameter φ becomes node dependent, and is a realisation of an Inverse Gamma distribution with parameters β_0 and β_1. Essentially, an increase in φ will mainly affect how prone the corresponding actor is to creating long-range connections, rather than short-range ones. This behaviour is in line with typical scenarios in large social networks, where hubs differ from ordinary nodes in that they entail connections between distant areas (or communities) of the graph, decreasing the average path length (Watts and Strogatz 1998).

We can approximate (3.16) and (3.17) and characterise the factorial moments of the degree of a random node as a function of the model parameters $\tau, \gamma, \beta_0, \beta_1$, allowing an assessment of the extent to which such models can represent heavy tails. Since the value of τ makes no difference here, we fix it to 1.

Table 2 shows that the variance of random effects does not have much influence on the average degree of the network. This is relevant for studying heavy tails, since sparser networks will naturally have a higher skewness index. Hence, if we keep the mean of the random effect constant and change the variance, not much of the change in the skewness index will be due to the network becoming sparser.

Figure 5.1 shows that an increase in the variance of the random effects does yield

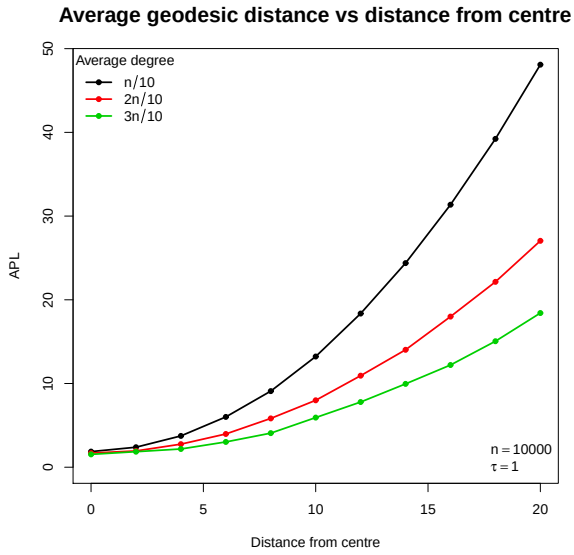

Figure 4.6: Average geodesic distance from a node as a function of its distance from the centre of the latent space. The network is composed of 10000 *nodes, with* $\tau = 1$*. Clearly, nodes which are closer to the centre will be better positioned to reach easily many other nodes, thus having a smaller APL index. Such heterogeneity in the connectivity structure characterises Gaussian LPMs and separates them from Erdős-Rényi random graphs, justifying the larger values for global APL.*

an increase in the skewness index, corresponding to a right-skewed heavy-tailed shape. Therefore, these two results indicate that the heaviness of the tails can be controlled by changing the variance of the random effects, without changing the average degree of the network by much. The smallest skewness index is obtained with a null variance for random effects, which corresponds to the Gaussian LPM.

But how heavy are the tails corresponding to a given positive skewness? Figure 5.2 shows the empirical degree frequencies obtained through simulations of Gaussian LPMREs. The two panels on the left side of Figure 5.2 show the degree distribution for a LPM (on both standard and log-log scale), where the variance of random effects is set to a very small value. The right-hand panels are obtained with the same parameters, except for the variance of the random effect, which is increased to 10^5. The average degrees for the two cases are: $0.151n$ and $0.144n$ respectively and the skewness indexes are -0.07 and 2.53 respectively. The log-log scale plots are represented to show that the decay switches from a high-order power-law (reasonably comparable to a Poissonian tail) to a power-law with an exponent which falls between 2 and 3.

The results shown confirm that random effects can extend the family of networks represented using LPMs. However, other features of interest are non-trivially influenced. Hence, we propose an empirical study to explore how random effects affect the asymptotic behaviour of LPMRE with respect to small-world behaviour and transitivity. Simulations

Table 2: Average degree of a network of 100 *actors for different values of mean and variance of the nodal random effects. The variance has very little impact on the overall average degree of the network. This is an important property which is needed to state that any increase of skewness is not due to the network getting sparser.*

	Variance						
Mean	0.0001	0.1	1	10	100	1000	100000
0.1	1.95	2.88	2.73	2.91	2.85	2.81	2.83
0.2	7.34	8.30	8.25	8.20	8.30	8.21	8.17
0.3	14.97	15.19	14.83	14.35	14.40	14.33	14.38
0.4	24.11	23.28	21.14	20.49	20.73	20.60	20.37

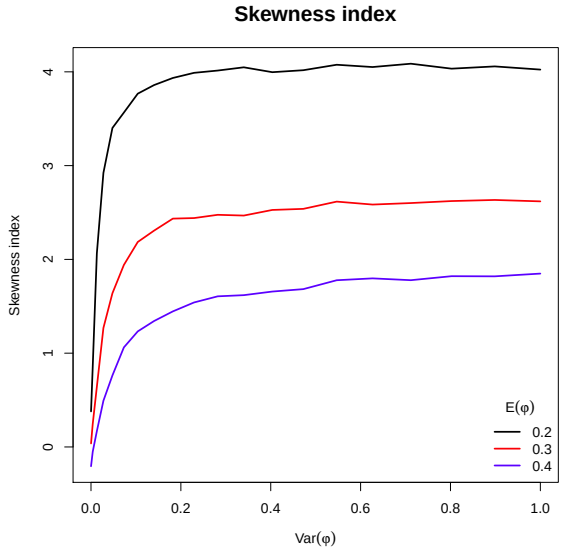

Figure 5.1: Skewness index versus variance of nodal random effects. An increase in the variance of the random effects leads to an increase of the skewness index, corresponding to heavier tails.

of LPMREs are very inefficient, so the results are rather limited. However, such a procedure is the only feasible one, since theoretical results on the LPMRE are not available. In fact, we are currently investigating alternative ways to approach this analysis using more rigorous theoretical frameworks.

In this experiment, we have selected a particular set of model parameters, generated a sequence of IID networks and studied the average features exhibited. Since we are interested in the asymptotic behaviour of APL and $\mathcal{C}$, we have held the average degree approximately constant by imposing $\gamma \propto n$, with n increasing. Figure 5.3 illustrates the results. The left panel shows that an increase in the variance of the random effects results in a smaller APL. Furthermore, the APL growth as a function of n becomes slower than the log function, exhibiting the small-world behaviour. The right panel represents instead

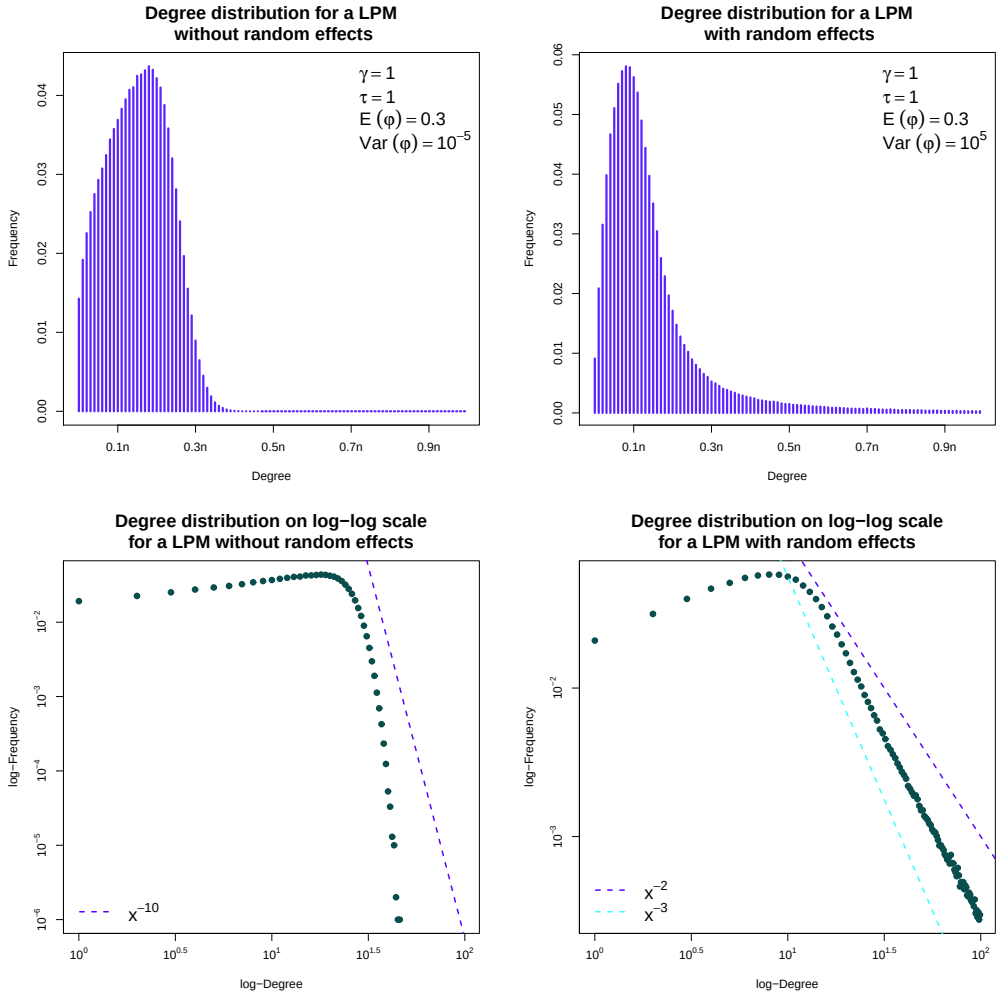

Figure 5.2: ***Top****: degree distributions for Gaussian LPMREs with null-variance random effects (****left****) and large-variance random effects (****right****).* ***Bottom****: corresponding degree distribution on the log-log scale. An increase in the variance of the random effects results in a heavier power-law tailed degree distribution. The average degrees are:* $0.151n$ *and* $0.144n$ *for the case on left and right respectively, while skewness indexes are* -0.07 *and* 2.53 *respectively.*

the empirical asymptotic clustering coefficient. Here, it appears that $\mathcal{C}$ tends to stabilise to a non-zero limiting value, which clearly depends on the variance of the random effects. Such interaction between the presence of hubs and the clustering coefficient could be somehow expected, since for an extreme case, the n-nodes star, $\mathcal{C}$ is equal to zero.

Considering the results shown in this Section, random effects can be regarded as a useful addition to LPMs to capture several important features that arise in large social networks.

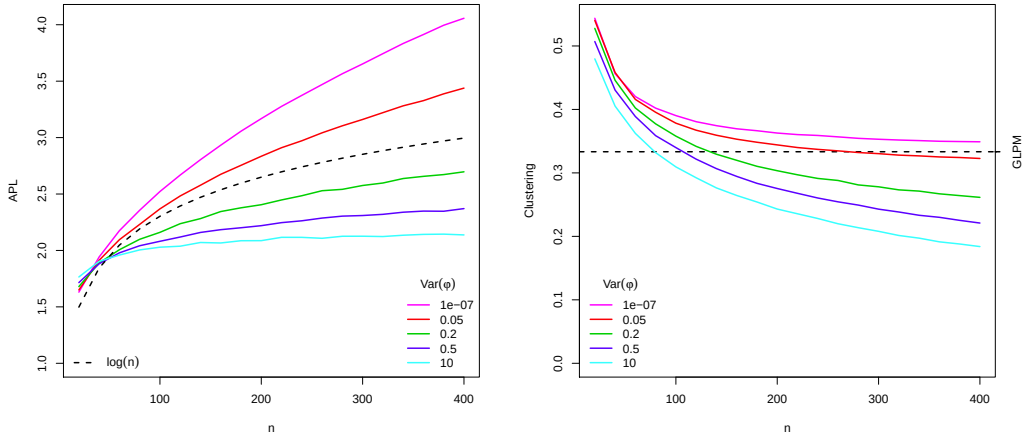

*Figure 5.3: APL (**left**) and clustering coefficient (**right**) as a function of n, holding an approximately constant average degree. The remaining model parameters are $\tau = 1$, $\mathbb{E}[\varphi] = 0.6$ and $\gamma = 0.05(n-1)$. The number of networks generated for each value of n is 1000. The dashed black lines represent the log function and the asymptotic value for $\mathcal{C}$ under the Gaussian LPM for the left and right panel respectively.*

6 Real data examples

We have characterised the models introduced by showing how some important statistics of realised networks depend on the parameters of LPMs. We now show that several well known real social networks have statistics that can be well captured by a fitted LPM, using the following datasets:

- **Dolphins**: This is a social network of frequent associations between 62 dolphins in a community living off Doubtful Sound, New Zealand (Lusseau et al. 2003).
- **Monks**: This describes the interpersonal relations among 18 monks in a monastery (Sampson 1968).
- **Florentine**: This describes the connections by marriage between 16 noble families in Florence during Renaissance (Padgett 1994).
- **Prison**: Data collected in the 1950s by John Gagnon from 67 prison inmates, each one being asked to specify his preferences among other participants (MacRae 1960).
- **High-tech**: This network contains the friendship ties among 36 employees of a hi-tech company, which were gathered by means of the question: who do you consider to be a personal friend? (Krackhardt 1999).
- **Math method**: 38 school superintendents were asked to indicate their friendship ties with other superintendents in the county with the following question: among the chief school administrators in Allegheny County (PA, USA), who are your three best friends? (Carlson 1965).

- **Sawmill**: 36 employees of a sawmill were asked to quantify the time they spent discussing work matters with each of their colleagues (Michael and Massey 1997).
- **San Juan**: Study carried out in a rural area in Costa Rica. Edges represent visiting frequencies between 75 families living in farms in a neighbourhood called San Juan Sur (De Nooy et al. 2011).
- **Network sciences** (1589 nodes): Coauthorship network of scientists working on network theory and experiment (Newman 2006).
- **Geometry** (7343 nodes): Coauthorship network of scientists working on computational geometry (Jones 2002).
- **Condensed Matter** (16726 nodes): Coauthorships between scientists posting preprints on the Condensed Matter E-Print Archive (Newman 2001).
- **High energy** (27770 nodes): Coauthorships between scientists posting preprints on the High-Energy Theory E-Print Archive (Newman 2001).

Where necessary, the datasets have been transformed into binary undirected (no self-edges) graphs, using standard reasonable procedures.

We can obtain the following network statistics for the Gaussian LPM using Theorem 1: the average degree $\bar{k}$, the clustering coefficient $\mathcal{C}$, the average path length APL and the skewness index S. Table 3 shows their observed and theoretical values for the smaller datasets.

The theoretical values shown in Table 3 correspond to model parameters chosen to match the observed with the theoretical $\bar{k}$ and $\mathcal{C}$. This simple criterion performs well for the networks presented, as indicated by Figure 6.1, which shows theoretical and observed degree distributions.

A slightly different study was carried out for the larger datasets, to assess to what extent the Gaussian LPMRE can represent the asymptotic scale-free decay of the degree distribution, for different orders of the power-law. We consider several collaboration networks where nodes correspond to authors and two nodes are linked if the corresponding scientists published a paper as coauthors. All the networks shown exhibit a power-law degree distribution, with different slopes, which vary in the range 1 to 4. Figure 6.2 shows the theoretical and observed degree distributions on the log-log scale, indicating that the asymptotic behaviour is reasonably well represented in all the cases.

7 Conclusions

The main contribution of this paper is to advance our understanding of Latent Position Models for networks by providing several probabilistic results. Our main results describe features of realised Latent Position networks, characterising their degree distribution,

Table 3: Theoretical and observed statistics for small-sized social networks. Statistics shown are the average degree $\bar{k}$, the clustering coefficient $\mathcal{C}$, the average path length APL and the skewness index S. Following the criterion described, the average degree and the clustering coefficient are matched exactly in every case, while the corresponding skewness index and average path length are fairly close to the observed counterparts.

Parameters		**Dolphins (n=62)**	k	$\mathcal{C}$	S	APL
τ	0.810	Observed	5.129	0.309	0.292	3.357
φ/γ	0.232	Theoretical	5.129	0.309	0.461	3.282

Parameters		**Monks (n=18)**	k	$\mathcal{C}$	S	APL
τ	0.763	Observed	6.667	0.465	0.877	1.68
φ/γ	2.115	Theoretical	6.667	0.465	-0.05	1.724

Parameters		**Florentine (n=16)**	k	$\mathcal{C}$	S	APL
τ	0.302	Observed	2.5	0.191	0.424	2.486
φ/γ	2.460	Theoretical	2.5	0.191	0.503	2.827

Parameters		**Prison (n=67)**	k	$\mathcal{C}$	S	APL
τ	0.776	Observed	4.239	0.288	0.855	3.355
φ/γ	0.180	Theoretical	4.239	0.288	0.562	3.831

Parameters		**High-tech (n=36)**	k	$\mathcal{C}$	S	APL
τ	0.913	Observed	5.056	0.372	0.785	2.360
φ/γ	0.376	Theoretical	5.056	0.372	0.376	2.749

Parameters		**Math method (n=38)**	k	$\mathcal{C}$	S	APL
τ	0.616	Observed	3.211	0.246	0.654	2.644
φ/γ	0.328	Theoretical	3.211	0.246	0.612	3.480

Parameters		**Sawmill (n=36)**	k	$\mathcal{C}$	S	APL
τ	0.550	Observed	3.444	0.230	2.290	3.138
φ/γ	0.436	Theoretical	3.444	0.230	0.558	3.210

Parameters		**San Juan (n=75)**	k	$\mathcal{C}$	S	APL
τ	0.657	Observed	4.133	0.245	1.622	3.485
φ/γ	0.186	Theoretical	4.133	0.245	0.579	3.883

the mixing properties of the degrees, the clustering coefficient and the path lengths' distribution. Although this work deals only with undirected graphs, the same results can be extended in a similar fashion to directed ones.

Gaussian LPMs have been shown not to be appropriate for modelling scale-free networks, since the average degree frequencies exhibit a left-skewed and truncated shape. However, modifying the basic LPM to include nodal random effects resulted in the ability of the model to represent power-law degree distributions of different slopes in both simulated and real networks.

It has been also shown that Gaussian LPMs have an asymptotically strictly positive clustering coefficient, in contrast to other well known models, such as Erdős-Rényi and Exponential Random Graph models, whose clustering coefficient is asymptotically zero.

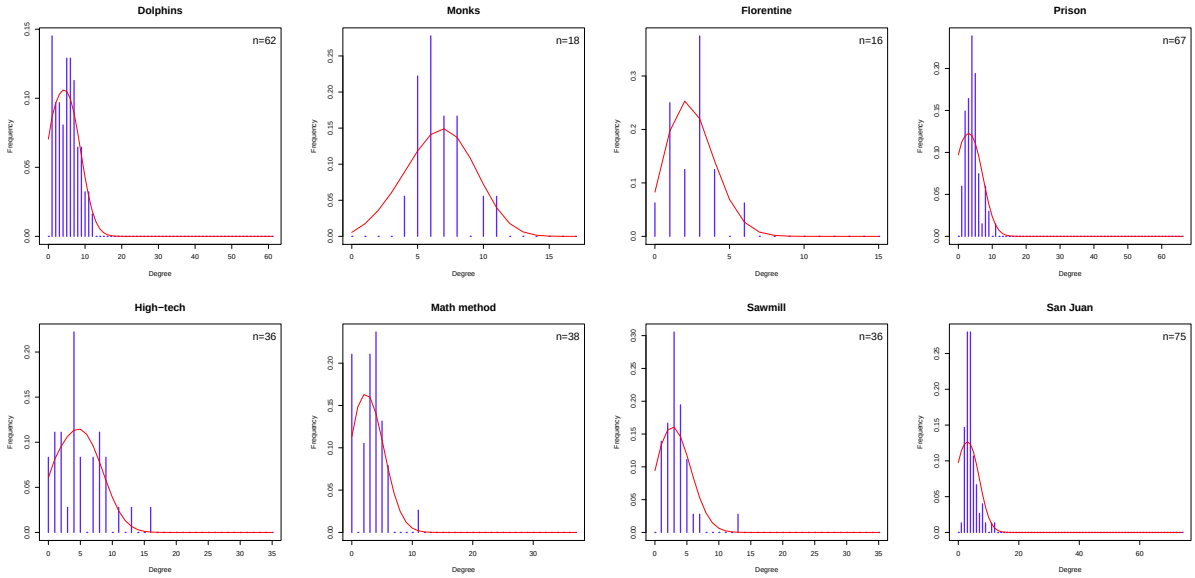

Figure 6.1: Comparison between the observed degree distributions (blue bars) and the theoretical ones (red lines) for several small-size real social networks. Datasets used (from top left by row): Dolphins, Monks, Florentine, Prison, High-tech, Math method, Sawmill, San Juan.

This result suggests that LPMs can generate highly clustered networks and that they can capture the persistent clustering behaviour of large social networks.

The average degree of the closest neighbours to a node has been characterised, showing that positive degree correlations arise in LPM networks. This is in line with observed social networks, where assortative mixing in the nodal degrees frequently occurs.

It has also been shown how the distribution of geodesic distances can be efficiently approximated, yielding an analysis of the asymptotic behaviour of the average path length. It appears that dense LPM networks have the same behaviour of Erdős-Rényi random graphs, while sparser LPM networks do not exhibit the small-world effect.

Through simulations, important advantages of using nodal random effects have been outlined, suggesting that the Gaussian LPMRE has properties that makes it suitable for modelling large social networks. An important extension of this work would be to develop new strategies to study analytically the LPMRE and LPCM.

Acknowledgements

The Insight Centre for Data Analytics is supported by Science Foundation Ireland under Grant Number SFI/12/RC/2289. Nial Friel and Riccardo Rastelli's research was also supported by a Science Foundation Ireland grant: 12/IP/1424. Adrian Raftery's research was supported by the Eunice Kennedy Shriver National Institute of Child Health and Development through NIH grants nos. R01 HD054511 and R01 HD070936, by Science

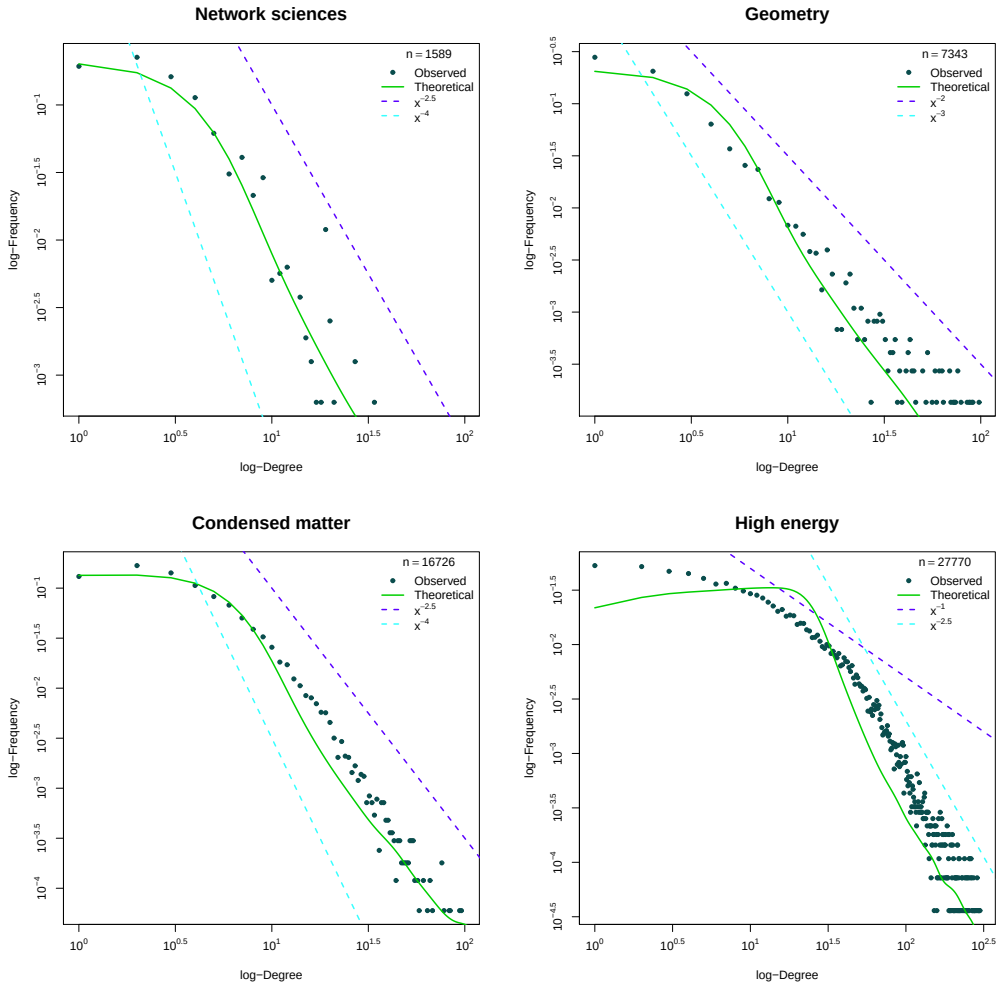

Figure 6.2: Empirical (blue dots) and theoretical (green line) degree distributions on log-log scale for various large citation networks. The datasets exhibit different asymptotic power-law orders. Gaussian LPMREs reasonably represent the asymptotic tendency of the degree distributions in every case. Datasets used: Network sciences (top left), Geometry (top right), Condensed matter (bottom left), High energy (bottom right).

Foundation Ireland grant 11/W.1/I2079 and by National Institutes of Health grant U54-HL127624.

A Appendix: proofs

A.1 Theorem 1

D1. This is straightforward since $\forall \mathbf{z}_s \in \mathbb{R}^d$:

$$\theta(\mathbf{z}_s) = Pr(y_{sj} = 1|\mathbf{z}_s) = \int_{\mathcal{Z}} p(\mathbf{z}_j)\, r(\mathbf{z}_s, \mathbf{z}_j)\, d\mathbf{z}_j. \tag{A.1}$$

D2.

$$\begin{aligned}
G(x) &= \sum_{k=0}^{n-1} x^k p_k = \sum_{k=0}^{n-1} x^k Pr(D_s = k) \\
&= \sum_{k=0}^{n-1} x^k \int_{\mathcal{Z}} \cdots \int_{\mathcal{Z}} p(\mathbf{z}_1) \cdots p(\mathbf{z}_n)\, Pr(D_s = k|P)\, d\mathbf{z}_1 \cdots d\mathbf{z}_n \\
&= \int_{\mathcal{Z}} \cdots \int_{\mathcal{Z}} \left[\prod_{j=1}^{n} p(\mathbf{z}_j)\right] \mathbb{E}\left[x^{D_s}\middle|P\right] d\mathbf{z}_1 \cdots d\mathbf{z}_n \\
&= \int_{\mathcal{Z}} \cdots \int_{\mathcal{Z}} \left[\prod_{j=1}^{n} p(\mathbf{z}_j)\right] \left\{\prod_{j=1}^{n} \mathbb{E}\left[x^{Y_{sj}}\middle|P\right]\right\} d\mathbf{z}_1 \cdots d\mathbf{z}_n \\
&= \int_{\mathcal{Z}} \cdots \int_{\mathcal{Z}} \left\{\prod_{j=1}^{n} p(\mathbf{z}_j)\left[x r(\mathbf{z}_s, \mathbf{z}_j) + 1 - r(\mathbf{z}_s, \mathbf{z}_j)\right]\right\} d\mathbf{z}_1 \cdots d\mathbf{z}_n \\
&= \int_{\mathcal{Z}} p(\mathbf{z}_s) \left\{\int_{\mathcal{Z}} p(\mathbf{z}_j)\left[x r(\mathbf{z}_s, \mathbf{z}_j) + 1 - r(\mathbf{z}_s, \mathbf{z}_j)\right] d\mathbf{z}_j\right\}^{n-1} d\mathbf{z}_s \\
&= \int_{\mathcal{Z}} p(\mathbf{z}_s) \left\{x \int_{\mathcal{Z}} p(\mathbf{z}_j)\, r(\mathbf{z}_s, \mathbf{z}_j)\, d\mathbf{z}_j + 1 - \int_{\mathcal{Z}} p(\mathbf{z}_j)\, r(\mathbf{z}_s, \mathbf{z}_j)\, d\mathbf{z}_j\right\}^{n-1} d\mathbf{z}_s \\
&= \int_{\mathcal{Z}} p(\mathbf{z}_s) \left[x\theta(\mathbf{z}_s) + 1 - \theta(\mathbf{z}_s)\right]^{n-1} d\mathbf{z}_s.
\end{aligned} \tag{A.2}$$

D3. The r-th factorial moment of D_s corresponds to the r-th derivative of G evaluated in 1:

$$\begin{aligned}
\frac{\partial^r G}{\partial x^r}(x) &= \int_{\mathcal{Z}} p(\mathbf{z}_s) \frac{\partial^r}{\partial x^r} \left[x\theta(\mathbf{z}_s) + 1 - \theta(\mathbf{z}_s)\right]^{n-1} d\mathbf{z}_s \\
&= \int_{\mathcal{Z}} p(\mathbf{z}_s)\, (n-1) \cdots (n-r)\, \theta(\mathbf{z}_s)^r \left[x\theta(\mathbf{z}_s) + 1 - \theta(\mathbf{z}_s)\right]^{n-r-1} d\mathbf{z}_s \\
&= \frac{(n-1)!}{(n-r-1)!} \int_{\mathcal{Z}} p(\mathbf{z}_s)\, \theta(\mathbf{z}_s)^r \left[x\theta(\mathbf{z}_s) + 1 - \theta(\mathbf{z}_s)\right]^{n-r-1} d\mathbf{z}_s;
\end{aligned} \tag{A.3}$$

and the final formula evaluated in $x = 1$ gives (3.3).

D4. The average degree is the first factorial moment, thus:

$$\bar{k} = G'(1) = \frac{(n-1)!}{(n-2)!} \int_{\mathcal{Z}} p(\mathbf{z}_s)\, \theta(\mathbf{z}_s)\, d\mathbf{z}_s = (n-1) \int_{\mathcal{Z}} p(\mathbf{z}_s)\, \theta(\mathbf{z}_s)\, d\mathbf{z}_s. \tag{A.4}$$

D5. The distribution of the degree of a random node can be recovered by differentiating G as well. Indeed, using (A.3), for every k:

$$p_k = \frac{1}{k!}\frac{\partial^r G}{\partial x^r}(0) = \binom{n-1}{k}\int_{\mathcal{Z}} p\left(\mathbf{z}_s\right)\theta\left(\mathbf{z}_s\right)\left[1-\theta\left(\mathbf{z}_s\right)\right]^{n-k-1} d\mathbf{z}_s. \tag{A.5}$$

D6. Define the PGF for the degree of a random node once its latent information is fixed to $\mathbf{z}_s$:

$$\begin{aligned}
\tilde{G}\left(x;\mathbf{z}_s\right) &= \sum_{k=0}^{n-1} x^k Pr\left(D_s = k|\mathbf{z}_s\right) \\
&= \int_{\mathcal{Z}}\cdots\int_{\mathcal{Z}}\left[\prod_{\substack{j=1\\ j\neq s}}^{n} p\left(\mathbf{z}_j\right)\right]\mathbb{E}\left[x^{D_s}\middle|P\right] d\mathbf{z}_{-s} \\
&= \left\{\int_{\mathcal{Z}} p\left(\mathbf{z}_j\right)\left[xr\left(\mathbf{z}_s,\mathbf{z}_j\right)+1-r\left(\mathbf{z}_s,\mathbf{z}_j\right)\right] d\mathbf{z}_j\right\}^{n-1} \\
&= \left\{x\theta\left(\mathbf{z}_s\right)+1-\theta\left(\mathbf{z}_s\right)\right\}^{n-1};
\end{aligned} \tag{A.6}$$

which is simply the PGF of a binomial random variable with parameters $n-1$ and $\theta\left(\mathbf{z}_s\right)$. Hence its average degree is $\bar{k}\left(\mathbf{z}_s\right) = \left(n-1\right)\theta\left(\mathbf{z}_s\right)$. Note that $d\mathbf{z}_{-s} = \prod_{j\neq s} d\mathbf{z}_j$.

D7. We now write down the PGF for the degree of a random neighbour of a node located in $\mathbf{z}_s$.

$$\begin{aligned}
H\left(x;\mathbf{z}_s\right) &= \sum_{k=0}^{n-1} x^k Pr\left(D_j = k|y_{sj}=1,\mathbf{z}_s\right) \\
&= \int_{\mathcal{Z}} p\left(\mathbf{z}_j|y_{sj}=1,\mathbf{z}_s\right)\sum_{k=0}^{n-1} x^k Pr\left(D_j = k|y_{sj}=1,\mathbf{z}_s,\mathbf{z}_j\right) d\mathbf{z}_j \\
&= \int_{\mathcal{Z}} p\left(\mathbf{z}_j|y_{sj}=1,\mathbf{z}_s\right)\mathbb{E}\left[x^{D_j}\middle|y_{sj}=1,\mathbf{z}_s,\mathbf{z}_j\right] d\mathbf{z}_j.
\end{aligned} \tag{A.7}$$

Note that $\mathbb{E}\left[x^{D_j}\middle|y_{sj}=1,\mathbf{z}_s,\mathbf{z}_j\right]$ corresponds to the PGF for the so called excess degree (Newman et al. 2001), i.e. the degree of a node at one extreme of an edge picked at random. Hence, such PGF is equal to $\frac{x\tilde{G}'(x;\mathbf{z}_j)}{\tilde{G}(1;\mathbf{z}_j)}$, where $\tilde{G}$ has been defined in (A.6). Then:

$$\begin{aligned}
H\left(x;\mathbf{z}_s\right) &= \int_{\mathcal{Z}} p\left(\mathbf{z}_j|y_{sj}=1,\mathbf{z}_s\right)\frac{x\tilde{G}'(x;\mathbf{z}_j)}{\tilde{G}\left(1;\mathbf{z}_j\right)} d\mathbf{z}_j \\
&= \int_{\mathcal{Z}} \frac{Pr\left(y_{sj}=1|\mathbf{z}_j,\mathbf{z}_s\right)p\left(\mathbf{z}_j\right)}{Pr\left(y_{sj}=1|\mathbf{z}_s\right)}\left\{x\left[x\theta\left(\mathbf{z}_j+1-\theta\left(\mathbf{z}_j\right)\right)\right]^{n-2}\right\} d\mathbf{z}_j \\
&= \frac{1}{\theta\left(\mathbf{z}_s\right)}\int_{\mathcal{Z}} p\left(\mathbf{z}_j\right) r\left(\mathbf{z}_j,\mathbf{z}_s\right)\left\{x\left[x\theta\left(\mathbf{z}_j+1-\theta\left(\mathbf{z}_j\right)\right)\right]^{n-2}\right\} d\mathbf{z}_j.
\end{aligned} \tag{A.8}$$

Its average degree is then given by:

$$\begin{aligned}\bar{k}_{nn}(\mathbf{z}_s) = H'(1;\mathbf{z}_s) &= \frac{1}{\theta(\mathbf{z}_s)}\int_{\mathcal{Z}} p(\mathbf{z}_j)\, r(\mathbf{z}_j,\mathbf{z}_s)\left\{1+(n-2)\,\theta(\mathbf{z}_j)\right\} d\mathbf{z}_j \\ &= 1+\frac{(n-2)}{\theta(\mathbf{z}_s)}\int_{\mathcal{Z}} p(\mathbf{z}_j)\, r(\mathbf{z}_j,\mathbf{z}_s)\,\theta(\mathbf{z}_j)\, d\mathbf{z}_j.\end{aligned} \tag{A.9}$$

D8. The PGF for the degree of a neighbour of a node with degree k is given by:

$$\begin{aligned}\tilde{H}(x;k) &= \sum_{r=0}^{n-1} x^r Pr(D_j = r|D_s = k, y_{sj}=1) \\ &= \sum_{r=0}^{n-1} x^r \int_{\mathcal{Z}} p(\mathbf{z}_s|D_s=k)\, Pr(D_j=r|\mathbf{z}_s, y_{sj}=1)\, d\mathbf{z}_s \\ &= \frac{1}{p_k}\int_{\mathcal{Z}} p(\mathbf{z}_s)\, Pr(D_s=k|\mathbf{z}_s)\, H(x;\mathbf{z}_s)\, d\mathbf{z}_s \\ &= \frac{1}{p_k}\int_{\mathcal{Z}} p(\mathbf{z}_s)\left[\frac{\partial^k}{\partial x^k}\tilde{G}(0;\mathbf{z}_s)\right] H(x;\mathbf{z}_s)\, d\mathbf{z}_s \\ &= \frac{1}{p_k}\int_{\mathcal{Z}} p(\mathbf{z}_s)\binom{n-1}{k}\theta(\mathbf{z}_s)^k\left[1-\theta(\mathbf{z}_s)\right]^{n-k-1} H(x;\mathbf{z}_s)\, d\mathbf{z}_s;\end{aligned} \tag{A.10}$$

and its first derivative evaluated in $x=1$ yields:

$$\bar{k}_{nn}(k) = \frac{1}{p_k}\int_{\mathcal{Z}} p(\mathbf{z}_s)\binom{n-1}{k}\theta(\mathbf{z}_s)^k\left[1-\theta(\mathbf{z}_s)\right]^{n-k-1}\bar{k}_{nn}(\mathbf{z}_s)d\mathbf{z}_s. \tag{A.11}$$

A.1.1 Proof for Corollary 1

Recall that a convolution of two Gaussian densities is still a Gaussian density:

$$\int_{\mathbb{R}^d} f_d(\mathbf{z}_i;\boldsymbol{\mu}_1,\gamma_1)\, f_d(\mathbf{z}_j-\mathbf{z}_i;\boldsymbol{\mu}_2,\gamma_2)\, d\mathbf{z}_i = f_d(\mathbf{z}_j;\boldsymbol{\mu}_1+\boldsymbol{\mu}_2,\gamma_1+\gamma_2), \tag{A.12}$$

for every $\mathbf{z}_i, \mathbf{z}_j, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2$ in $\mathbb{R}^d$ and every positive real numbers γ_1 and γ_2.

That being said:

D1.

$$\begin{aligned}\theta(\mathbf{z}_s) &= \int_{\mathbb{R}^d} f_d(\mathbf{z}_j;\mathbf{0},\gamma)\,\tau\,(2\pi\varphi)^{\frac{d}{2}}\, f_d(\mathbf{z}_s-\mathbf{z}_j;\mathbf{0},\varphi)\, d\mathbf{z}_j \\ &= \tau\,(2\pi\varphi)^{\frac{d}{2}}\, f_d(\mathbf{z}_s;\mathbf{0},\gamma+\varphi) \\ &= \tau\left(\frac{\varphi}{\gamma+\varphi}\right)^{\frac{d}{2}}\exp\left\{-\frac{1}{2(\gamma+\varphi)}\mathbf{z}_s^t\mathbf{z}_s\right\}.\end{aligned} \tag{A.13}$$

D3.

$$\begin{aligned}
\frac{\partial^r G}{\partial x^r}(1) &= \frac{(n-1)!}{(n-r-1)!}\int_{\mathbb{R}^d} f_d\left(\mathbf{z}_s;\mathbf{0},\gamma\right)\theta(\mathbf{z}_s)^r d\mathbf{z}_s \\
&= \frac{(n-1)!}{(n-r-1)!}\tau^r\left(\frac{\varphi}{\gamma+\varphi}\right)^{\frac{rd}{2}}\int_{\mathbb{R}^d} f_d\left(\mathbf{z}_s;\mathbf{0},\gamma\right)\exp\left\{-\frac{r}{2\left(\gamma+\varphi\right)}\mathbf{z}_s^t\mathbf{z}_s\right\}d\mathbf{z}_s \\
&= \frac{(n-1)!}{(n-r-1)!}\tau^r\left(\frac{\varphi}{\gamma+\varphi}\right)^{\frac{rd}{2}}\left\{2\pi\frac{(\gamma+\varphi)}{r}\right\}^{\frac{d}{2}}\times \\
&\qquad\qquad \times\int_{\mathbb{R}^d} f_d\left(\mathbf{z}_s;\mathbf{0},\gamma\right) f_d\left(\mathbf{z}_s;\mathbf{0},\frac{\gamma+\varphi}{r}\right)d\mathbf{z}_s \\
&= \frac{(n-1)!}{(n-r-1)!}\tau^r\left(\frac{\varphi}{\gamma+\varphi}\right)^{\frac{rd}{2}}\left\{2\pi\frac{(\gamma+\varphi)}{r}\right\}^{\frac{d}{2}}\left\{2\pi\frac{\left[(r+1)\gamma+\varphi\right]}{r}\right\}^{-\frac{d}{2}} \\
&= \frac{(n-1)!}{(n-r-1)!}\tau^r\left\{\frac{\varphi^r}{(\gamma+\varphi)^{r-1}\left[(r+1)\gamma+\varphi\right]}\right\}^{\frac{d}{2}}
\end{aligned} \tag{A.14}$$

D4.

$$\bar{k} = G'(1) = (n-1)\tau\left\{\frac{\varphi}{2\gamma+\varphi}\right\}^{\frac{d}{2}} \tag{A.15}$$

D7.

$$\begin{aligned}
\bar{k}_{nn}(\mathbf{z}_s) &= 1+\frac{(n-2)}{\theta\left(\mathbf{z}_s\right)}\int_{\mathbb{R}^d} p\left(\mathbf{z}_j\right) r\left(\mathbf{z}_s,\mathbf{z}_j\right)\theta\left(\mathbf{z}_j\right)d\mathbf{z}_j \\
&= 1+\frac{(n-2)}{\theta\left(\mathbf{z}_s\right)}\tau^2\left(2\pi\varphi\right)^d\times \\
&\qquad\qquad\times\int_{\mathbb{R}^d} f_d\left(\mathbf{z}_j;\mathbf{0},\gamma\right) f_d\left(\mathbf{z}_j;\mathbf{0},\gamma+\varphi\right) f_d\left(\mathbf{z}_s-\mathbf{z}_j;\mathbf{0},\varphi\right)d\mathbf{z}_j \\
&= 1+\frac{(n-2)}{\theta\left(\mathbf{z}_s\right)}\tau^2\left(2\pi\varphi\right)^d\left\{2\pi\left(2\gamma+\varphi\right)\right\}^{-\frac{d}{2}}\times \\
&\qquad\qquad\times\int_{\mathbb{R}^d} f_d\left(\mathbf{z}_j;\mathbf{0},\frac{\gamma\left(\gamma+\varphi\right)}{2\gamma+\varphi}\right) f_d\left(\mathbf{z}_s-\mathbf{z}_j;\mathbf{0},\varphi\right)d\mathbf{z}_j \\
&= 1+(n-2)\tau\left(\frac{\varphi}{2\gamma+\varphi}\right)^{\frac{d}{2}}\frac{f_d\left(\mathbf{z}_s;\mathbf{0},\varphi+\frac{\gamma(\gamma+\varphi)}{2\gamma+\varphi}\right)}{f_d\left(\mathbf{z}_s;\mathbf{0},\gamma+\varphi\right)} \\
&= 1+\bar{k}\left(\frac{n-2}{n-1}\right)\frac{f_d\left(\mathbf{z}_s;\mathbf{0},\frac{\gamma^2+3\gamma\varphi+\varphi^2}{2\gamma+\varphi}\right)}{f_d\left(\mathbf{z}_s;\mathbf{0},\gamma+\varphi\right)}.
\end{aligned} \tag{A.16}$$

A.1.2 Proof for Corollary 2

D1.

$$\begin{aligned}\theta(\mathbf{z}_s) &= \int_{\mathbb{R}^d} \sum_{g=1}^{G} \pi_g f_d\left(\mathbf{z}_j; \boldsymbol{\mu}_g, \gamma_g\right) \tau\left(2\pi\varphi\right)^{\frac{d}{2}} f_d\left(\mathbf{z}_s - \mathbf{z}_j; \mathbf{0}, \varphi\right) d\mathbf{z}_j \\ &= \tau\left(2\pi\varphi\right)^{\frac{d}{2}} \sum_{g=1}^{G} \pi_g \int_{\mathbb{R}^d} f_d\left(\mathbf{z}_j; \boldsymbol{\mu}_g, \gamma_g\right) f_d\left(\mathbf{z}_s - \mathbf{z}_j; \mathbf{0}, \varphi\right) d\mathbf{z}_j \\ &= \tau\left(2\pi\varphi\right)^{\frac{d}{2}} \sum_{g=1}^{G} \pi_g f_d\left(\mathbf{z}_s; \boldsymbol{\mu}_g, \gamma_g + \varphi\right).\end{aligned} \tag{A.17}$$

D4.

$$\begin{aligned}\bar{k} &= (n-1) \int_{\mathbb{R}^d} \sum_{g=1}^{G} \pi_g f_d\left(\mathbf{z}_s; \boldsymbol{\mu}_g, \gamma_g\right) \tau\left(2\pi\varphi\right)^{\frac{d}{2}} \sum_{h=1}^{G} \pi_h f_d\left(\mathbf{z}_s; \boldsymbol{\mu}_h, \gamma_h + \varphi\right) d\mathbf{z}_s \\ &= (n-1)\tau\left(2\pi\varphi\right)^{\frac{d}{2}} \sum_{g=1}^{G}\sum_{h=1}^{G} \pi_g\pi_h \int_{\mathbb{R}^d} f_d\left(\mathbf{z}_s; \boldsymbol{\mu}_g, \gamma_g\right) f_d\left(\mathbf{z}_s; \boldsymbol{\mu}_h, \gamma_h + \varphi\right) d\mathbf{z}_s \\ &= (n-1)\tau\left(2\pi\varphi\right)^{\frac{d}{2}} \sum_{g=1}^{G}\sum_{h=1}^{G} \pi_g\pi_h f_d\left(\boldsymbol{\mu}_g - \boldsymbol{\mu}_h; \mathbf{0}, \gamma_g + \gamma_h + \varphi\right).\end{aligned} \tag{A.18}$$

While **D7** is straightforward from (3.5).

A.2 Proof of Proposition 2

First, we recall a few properties of the Gaussian distribution through a Lemma:

Lemma 1. *Let $f_d\left(\cdot; \boldsymbol{\mu}, \gamma\right)$ denote the d-dimensional Gaussian density centred in $\boldsymbol{\mu}$, with covariance matrix $\gamma \boldsymbol{I}_d$. Let also $\boldsymbol{x}, \boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^d$ and $a, b, \alpha \in \mathbb{R}^+$. Then:*

$$f_d\left(\boldsymbol{x}; \boldsymbol{u}, a\right) f_d\left(\boldsymbol{x}; \boldsymbol{v}, b\right) = f_d\left(\boldsymbol{u} - \boldsymbol{v}; \boldsymbol{0}, a + b\right) f_d\left(\boldsymbol{x}; \frac{b\boldsymbol{u} + a\boldsymbol{v}}{a+b}, \frac{ab}{a+b}\right); \tag{A.19}$$

$$f_d\left(\alpha\boldsymbol{x}; \boldsymbol{u}, a\right) = \alpha^{-d} f_d\left(\boldsymbol{x}; \frac{\boldsymbol{u}}{\alpha}, \frac{a}{\alpha^2}\right). \tag{A.20}$$

Here follows the proof of Proposition 2 by mathematical induction on k. If $k = 1$, then:

$$I_1(\mathbf{z}_i, \mathbf{z}_j) = h_1 f_d\left(\mathbf{z}_j - \alpha_1\mathbf{z}_i; \mathbf{0}, \omega_1\right) = \tau\left(2\pi\varphi\right)^{\frac{d}{2}} f_d\left(\mathbf{z}_j - \mathbf{z}_i; \mathbf{0}, \varphi\right) = r\left(\mathbf{z}_i, \mathbf{z}_j\right). \tag{A.21}$$

Now assume that $I_k(\mathbf{z}_i, \mathbf{z}_j) = h_k f_d\left(\mathbf{z}_j - \alpha_k\mathbf{z}_i; \mathbf{0}, \omega_k\right)$, then we need to prove that

$$I_{k+1}(\mathbf{z}_i, \mathbf{z}_j) = h_{k+1} f_d\left(\mathbf{z}_j - \alpha_{k+1}\mathbf{z}_i; \mathbf{0}, \omega_{k+1}\right),$$

where $h_{k+1}, \alpha_{k+1}, \omega_{k+1}$ are defined recursively by (3.21).

$$\begin{aligned} I_{k+1}(\mathbf{z}_i, \mathbf{z}_j) &= \int_{\mathcal{Z}} \cdots \int_{\mathcal{Z}} p(\mathbf{z}_1) \dots p(\mathbf{z}_k) r(\mathbf{z}_i, \mathbf{z}_1) \cdots r(\mathbf{z}_k, \mathbf{z}_j) \, d\mathbf{z}_1 \cdots d\mathbf{z}_k \\ &= \int_{\mathcal{Z}} p(\mathbf{z}_k) \, r(\mathbf{z}_k, \mathbf{z}_j) \int_{\mathcal{Z}} \cdots \int_{\mathcal{Z}} p(\mathbf{z}_1) \dots p(\mathbf{z}_{k-1}) \times \\ &\quad \times r(\mathbf{z}_i, \mathbf{z}_1) \cdots r(\mathbf{z}_{k-1}, \mathbf{z}_k) \, d\mathbf{z}_1 \cdots d\mathbf{z}_k \\ &= \int_{\mathcal{Z}} p(\mathbf{z}_k) \, r(\mathbf{z}_k, \mathbf{z}_j) \, I_k(\mathbf{z}_i, \mathbf{z}_k) \, d\mathbf{z}_k \\ &= \int_{\mathcal{Z}} p(\mathbf{x}) \, r(\mathbf{x}, \mathbf{z}_j) \, I_k(\mathbf{z}_i, \mathbf{x}) \, d\mathbf{x}. \end{aligned} \tag{A.22}$$

Now, we introduce the Gaussian LPM assumptions and use the results of the Lemma 1:

$$\begin{aligned} I_{k+1}(\mathbf{z}_i, \mathbf{z}_j) &= \tau (2\pi\varphi)^{\frac{d}{2}} h_k \int_{\mathbb{R}^d} f_d(\mathbf{x}; \mathbf{0}, \gamma) \, f_d(\mathbf{x} - \mathbf{z}_j; \mathbf{0}, \varphi) \, f_d(\mathbf{x} - \alpha_k \mathbf{z}_i; \mathbf{0}, \omega_k) \, d\mathbf{x} \\ &= \tau (2\pi\varphi)^{\frac{d}{2}} h_k \times \\ &\quad \times \int_{\mathbb{R}^d} f_d(\mathbf{x} - \mathbf{z}_j; \mathbf{0}, \varphi) \, f_d(-\alpha_k \mathbf{z}_i; \mathbf{0}, \omega_k + \gamma) \, f_d\left(\mathbf{x}; \frac{\gamma \alpha_k \mathbf{z}_i}{\omega_k + \gamma}, \frac{\omega_k \gamma}{\omega_k + \gamma}\right) d\mathbf{x} \\ &= \tau (2\pi\varphi)^{\frac{d}{2}} h_k \alpha^{-d} f_d\left(\mathbf{z}_i; \mathbf{0}, \frac{\omega_k + \gamma}{\alpha_k^2}\right) \times \\ &\quad \times \int_{\mathbb{R}^d} f_d(\mathbf{x} - \mathbf{z}_j; \mathbf{0}, \varphi) \, f_d\left(\mathbf{x}; \frac{\gamma \alpha_k \mathbf{z}_i}{\omega_k + \gamma}, \frac{\omega_k \gamma}{\omega_k + \gamma}\right) d\mathbf{x} \\ &= h_{k+1} f_d\left(\mathbf{z}_j; \frac{\gamma \alpha_k \mathbf{z}_i}{\omega_k + \gamma}, \frac{\omega_k \gamma + \omega_k \varphi + \varphi \gamma}{\omega_k + \gamma}\right) \\ &= h_{k+1} f_d(\mathbf{z}_j - \alpha_{k+1} \mathbf{z}_i; \mathbf{0}.\omega_{k+1}). \end{aligned} \tag{A.23}$$

A.3 Proof of Corollary 3

Let G be the PGF of the random variable D, denoting the degree of a node picked at random. Then the r-th derivative of G evaluated in 1 is equal to the r-th factorial moment of D, denoted here c_r:

$$c_r = \frac{\partial^r G}{\partial x^r}(1) = \mathbb{E}[D(D-1)\cdots(D-r+1)]. \tag{A.24}$$

In particular:

$$c_1 = \mathbb{E}[D] = m_1 \tag{A.25}$$

$$c_2 = \mathbb{E}[D(D-1)] = \mathbb{E}\left[D^2\right] - \mathbb{E}[D] = m_2 - m_1 \tag{A.26}$$

$$\implies m_2 = c_1 + c_2, \tag{A.27}$$

where m_1 and m_2 denote the first two non-central moments of D. That being said, using Corollary 1 the dispersion index can be evaluated exactly:

$$\begin{aligned}
\mathcal{D} &= \frac{\mathbb{E}\left[(D-m_1)^2\right]}{m_1} = \frac{m_2 - m_1^2}{m_1} = \frac{m_2}{m_1} - m_1 = 1 + \frac{c_2}{c_1} - c_1 \\
&= 1 + \frac{(n-1)(n-2)\tau^2 \left\{\frac{\varphi^2}{(\gamma+\varphi)(3\gamma+\varphi)}\right\}^{\frac{d}{2}}}{(n-1)\tau\left\{\frac{\varphi}{2\gamma+\varphi}\right\}^{\frac{d}{2}}} - (n-1)\tau\left\{\frac{\varphi}{2\gamma+\varphi}\right\}^{\frac{d}{2}} \\
&= 1 + (n-2)\tau\left\{\frac{\varphi(2\gamma+\varphi)}{(\gamma+\varphi)(3\gamma+\varphi)}\right\}^{\frac{d}{2}} - (n-1)\tau\left\{\frac{\varphi}{2\gamma+\varphi}\right\}^{\frac{d}{2}},
\end{aligned} \tag{A.28}$$

which proves the corollary. Also, when $d = 2$, the threshold between underdispersion and overdispersion is given by:

$$\frac{(n-2)(2\gamma+\varphi)}{(\gamma+\varphi)(3\gamma+\varphi)} - \frac{(n-1)}{(2\gamma+\varphi)} = 0. \tag{A.29}$$

Now, recalling that $\varphi > 0$ and $\gamma > 0$, this is equivalent to:

$$\begin{aligned}
&(n-2)(2\gamma+\varphi)^2 - (n-1)(\gamma+\varphi)(3\gamma+\varphi) = 0 \\
\Rightarrow &\varphi^2 + 4\gamma\varphi + 5\gamma^2 - n\gamma^2 = 0 \\
\Rightarrow &\varphi = \gamma\left(-2 \pm \sqrt{n-1}\right).
\end{aligned} \tag{A.30}$$

One solution is negative thus not feasible, then the threshold is given by:

$$\varphi = \gamma\left(\sqrt{n-1} - 2\right).$$

A.4 Proof of Proposition 1

Formula in (3.18) is straightforward since it is obtained by conditioning on the latent information. We now show how to obtain the exact formula (3.19) under the Gaussian LPM. We solve the numerator $\mathcal{C}_N$ and the denominator $\mathcal{C}_D$ independently.

$$\begin{aligned}
\mathcal{C}_D &= \int_{\mathbb{R}^d}\int_{\mathbb{R}^d}\int_{\mathbb{R}^d} p(\mathbf{z}_i)p(\mathbf{z}_k)p(\mathbf{z}_j) r(\mathbf{z}_i,\mathbf{z}_k)\, r(\mathbf{z}_k,\mathbf{z}_j)\, d\mathbf{z}_k d\mathbf{z}_i d\mathbf{z}_j \\
&= \int_{\mathbb{R}^d} p(\mathbf{z}_k)\left\{\int_{\mathbb{R}^d} p(\mathbf{z}_i) r(\mathbf{z}_i,\mathbf{z}_k)\, d\mathbf{z}_i\right\}\left\{\int_{\mathbb{R}^d} p(\mathbf{z}_j) r(\mathbf{z}_k,\mathbf{z}_j)\, d\mathbf{z}_j\right\} d\mathbf{z}_k \\
&= \int_{\mathbb{R}^d} p(\mathbf{z}_k)\theta(\mathbf{z}_k)^2\, d\mathbf{z}_k \\
&= \frac{G''(1)}{(n-1)(n-2)} \\
&= \tau^2\left\{\frac{\varphi^2}{(\gamma+\varphi)(3\gamma+\varphi)}\right\}^{\frac{d}{2}}
\end{aligned} \tag{A.31}$$

Now we solve the numerator.

$$\begin{aligned}
\mathcal{C}_N &= \int_{\mathbb{R}^d}\int_{\mathbb{R}^d}\int_{\mathbb{R}^d} p(\mathbf{z}_i)p(\mathbf{z}_k)p(\mathbf{z}_j)r\left(\mathbf{z}_i,\mathbf{z}_k\right)r\left(\mathbf{z}_k,\mathbf{z}_j\right)r\left(\mathbf{z}_j,\mathbf{z}_i\right)d\mathbf{z}_i d\mathbf{z}_k d\mathbf{z}_j \\
&= \int_{\mathbb{R}^d} p(\mathbf{z}_i)\int_{\mathbb{R}^d} p(\mathbf{z}_j)r\left(\mathbf{z}_j,\mathbf{z}_i\right)\left\{\int_{\mathbb{R}^d} p(\mathbf{z}_k)r\left(\mathbf{z}_i,\mathbf{z}_k\right)r\left(\mathbf{z}_k,\mathbf{z}_j\right)d\mathbf{z}_k\right\}d\mathbf{z}_j d\mathbf{z}_i \\
&= \int_{\mathbb{R}^d} p(\mathbf{z}_i)\int_{\mathbb{R}^d} p(\mathbf{z}_j)r\left(\mathbf{z}_j,\mathbf{z}_i\right)I_2\left(\mathbf{z}_i,\mathbf{z}_j\right)d\mathbf{z}_j d\mathbf{z}_i \\
&= \int_{\mathbb{R}^d} p(\mathbf{z}_i)I_3(\mathbf{z}_i,\mathbf{z}_i)d\mathbf{z}_i
\end{aligned} \tag{A.32}$$

where $I_k(\mathbf{z}_i,\mathbf{z}_j)$ is defined in 3.20 for every $k\in\mathbb{N}^0$, $\mathbf{z}_i\in\mathbb{R}^d$ and $\mathbf{z}_j\in\mathbb{R}^d$.

For more clarity, we define the recurring quantity

$$\lambda = \varphi^2 + 3\gamma\varphi + \gamma^2. \tag{A.33}$$

We first discover the quantities needed to write $I_3(\mathbf{z}_i,\mathbf{z}_i)$ explicitly:

$$\begin{cases} \alpha_1 &= 1 \\ \omega_1 &= \varphi \\ h_1 &= \tau\left(2\pi\varphi\right)^{\frac{d}{2}} \end{cases}; \qquad \begin{cases} \alpha_2 &= \frac{\gamma}{\gamma+\varphi} \\ \omega_2 &= \frac{\varphi(2\gamma+\varphi)}{\gamma+\varphi} \\ h_2 &= \tau^2\left(2\pi\varphi\right)^d f_d\left(\mathbf{z}_i;\mathbf{0},\gamma+\varphi\right) \end{cases}; \tag{A.34}$$

$$\alpha_3 = \frac{\alpha_2\gamma}{\omega_2+\gamma} = \frac{\gamma^2}{\lambda}; \tag{A.35}$$

$$\omega_3 = \frac{\omega_2\varphi+\omega_2\gamma+\gamma\varphi}{\omega_2+\gamma} = \frac{\varphi\left(\gamma+\varphi\right)\left(3\gamma+\varphi\right)}{\lambda}; \tag{A.36}$$

$$h_3 = \tau^3\left(2\pi\varphi\right)^{\frac{3}{2}d} f_d\left(\mathbf{z}_i;\mathbf{0},\gamma+\varphi\right)\left(\frac{\gamma+\varphi}{\gamma}\right)^d f_d\left(\mathbf{z}_i;\mathbf{0},\frac{\lambda\left(\gamma+\varphi\right)}{\gamma^2}\right). \tag{A.37}$$

Now, for h_3, we use Lemma 1 and join the two Gaussian densities:

$$\begin{aligned}
h_3 &= \tau^3\left(2\pi\varphi\right)^{\frac{3}{2}d}\left(\frac{\gamma+\varphi}{\gamma}\right)^d\left\{2\pi\frac{\left(\gamma+\varphi\right)^2\left(2\gamma+\varphi\right)}{\gamma^2}\right\}^{-\frac{d}{2}} f_d\left(\mathbf{z}_i;\mathbf{0},\frac{\lambda}{2\gamma+\varphi}\right) \\
&= \tau^3\left(2\pi\varphi\right)^d\left\{\frac{\varphi}{2\gamma+\varphi}\right\}^{\frac{d}{2}} f_d\left(\mathbf{z}_i;\mathbf{0},\frac{\lambda}{2\gamma+\varphi}\right).
\end{aligned} \tag{A.38}$$

Also:

$$\left(1-\alpha_3\right) = \frac{\varphi\left(3\gamma+\varphi\right)}{\lambda} \tag{A.39}$$

$$\frac{\omega_3}{\left(1-\alpha_3\right)^2} = \frac{\lambda\left(\gamma+\varphi\right)}{\varphi\left(3\gamma+\varphi\right)} \tag{A.40}$$

$$\tag{A.41}$$

Then, it follows:

$$\begin{aligned} I_3(\mathbf{z}_i, \mathbf{z}_i) &= h_3 \left(1-\alpha_3\right)^{-d} f_d\left(\mathbf{z}_i; \mathbf{0}, \frac{\omega_3}{\left(1-\alpha_3\right)^2}\right) \\ &= \tau^3 \left(2\pi\varphi\right)^d \left\{\frac{\varphi}{2\gamma+\varphi}\right\}^{\frac{d}{2}} f_d\left(\mathbf{z}_i; \mathbf{0}, \frac{\lambda}{2\gamma+\varphi}\right) \times \\ &\quad \times \left\{\frac{\lambda}{\varphi\left(3\gamma+\varphi\right)}\right\}^d f_d\left(\mathbf{z}_i; \mathbf{0}, \frac{\lambda\left(\gamma+\varphi\right)}{\varphi\left(3\gamma+\varphi\right)}\right). \end{aligned} \tag{A.42}$$

Collapsing again the Gaussian densities:

$$I_3(\mathbf{z}_i, \mathbf{z}_i) = \tau^3 \left\{\frac{2\pi\varphi^2}{2\left(3\gamma+\varphi\right)}\right\}^{\frac{d}{2}} f_d\left(\mathbf{z}_i; \mathbf{0}, \frac{\gamma+\varphi}{2}\right) \tag{A.43}$$

We can now obtain the final result for the numerator:

$$\begin{aligned} \mathcal{C}_N &= \int_{\mathbb{R}^d} p(\mathbf{z}_i) I_3(\mathbf{z}_i, \mathbf{z}_i) d\mathbf{z}_i \\ &= \tau^3 \left\{\frac{2\pi\varphi^2}{2\left(3\gamma+\varphi\right)}\right\}^{\frac{d}{2}} \int_{\mathbb{R}^d} f_d\left(\mathbf{z}_i; \mathbf{0}, \gamma\right) f_d\left(\mathbf{z}_i; \mathbf{0}, \frac{\gamma+\varphi}{2}\right) d\mathbf{z}_i \\ &= \tau^3 \left\{\frac{\varphi^2}{\left(3\gamma+\varphi\right)^2}\right\}^{\frac{d}{2}} \end{aligned} \tag{A.44}$$

The final formula for the clustering coefficient follows:

$$\mathcal{C} = \frac{\mathcal{C}_N}{\mathcal{C}_D} = \frac{\tau^3 \left\{\frac{\varphi^2}{(3\gamma+\varphi)^2}\right\}^{\frac{d}{2}}}{\tau^2 \left\{\frac{\varphi^2}{(\gamma+\varphi)(3\gamma+\varphi)}\right\}^{\frac{d}{2}}} = \tau \left(\frac{\gamma+\varphi}{3\gamma+\varphi}\right)^{\frac{d}{2}}. \tag{A.45}$$

References

Airoldi, E. M., D. M. Blei, S. E. Fienberg, and E. P. Xing. 2008. Mixed membership stochastic blockmodels. In *Journal of machine learning research*, 1981–2014. Vol. 9.

Albert, R., H. Jeong, and A. L. Barabási. 2000. Error and attack tolerance of complex networks. *Nature* 406 (6794): 378–382.

Albert, R., H. Jeong, and A. L. Barabási. 1999. Internet: diameter of the world-wide web. *Nature* 401 (6749): 130–131.

Amaral, L. A. N., A. Scala, M. Barthelemy, and H. E. Stanley. 2000. Classes of small-world networks. *Proceedings of the national academy of sciences* 97 (21): 11149–11152.

Ambroise, C., and C. Matias. 2012. New consistent and asymptotically normal parameter estimates for random-graph mixture models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 74 (1): 3–35.

Barabási, A. L., and R. Albert. 1999. Emergence of scaling in random networks. *Science* 286 (5439): 509–512.

Boguná, M., and R. Pastor-Satorras. 2003. Class of correlated random networks with hidden variables. *Physical Review E* 68 (3): 036112.

Caimo, A., and N. Friel. 2011. Bayesian inference for exponential random graph models. *Social Networks* 33 (1): 41–55.

Caldarelli, G., A. Capocci, P. De Los Rios, and M. A. Muñoz. 2002. Scale-free networks from varying vertex intrinsic fitness. *Physical review letters* 89 (25): 258702.

Cao, X., and M. D. Ward. 2014. Do democracies attract portfolio investment? transnational portfolio investments modeled as dynamic network. *International Interactions* 40:216–245.

Carlson, R. O. 1965. *Adoption of educational innovations*. ERIC.

Channarond, A., J. J. Daudin, and S. Robin. 2012. Classification and estimation in the stochastic blockmodel based on the empirical degrees. *Electronic Journal of Statistics* 6:2574–2601.

Chatterjee, S., and P. Diaconis. 2013. Estimating and understanding exponential random graph models. *The Annals of Statistics* 41 (5): 2428–2461.

Chiu, G. S., and A. H. Westveld. 2014. A statistical social network model for consumption data in trophic food webs. *Statistical Methodology* 17:139–160.

Chiu, G. S., and A. H. Westveld. 2011. A unifying approach for food webs, phylogeny, social networks, and statistics. *Proceedings of the National Academy of Sciences* 108:15881–15886.

Daudin, J. J., F. Picard, and S. Robin. 2008. A mixture model for random graphs. *Statistics and computing* 18 (2): 173–183.

De Nooy, W., A. Mrvar, and V. Batagelj. 2011. *Exploratory social network analysis with pajek*. Vol. 27. Cambridge University Press.

Deprez, P., and M. V. Wüthrich. 2013. Poisson heterogeneous random-connection model. *arXiv:1312.1948*.

Dunbar, R. I. M. 1992. Neocortex size as a constraint on group size in primates. *Journal of Human Evolution* 22 (6): 469–493.

Frank, O., and D. Strauss. 1986. Markov graphs. *Journal of the american Statistical association* 81 (395): 832–842.

Fronczak, A., P. Fronczak, and J. A. Hołyst. 2004. Average path length in random networks. *Physical Review E* 70 (5): 056110.

Gollini, I., and T. B. Murphy. 2014. Joint modelling of multiple network views. *Journal of Computational and Graphical Statistics*.

Handcock, M. S., A. E. Raftery, and J. M. Tantrum. 2007. Model-based clustering for social networks. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 170 (2): 301–354.

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

Jones, B. 2002. *Computational geometry database*.

Kiss, I. Z., and D. M. Green. 2008. Comment on "properties of highly clustered networks". *Physical Review E* 78 (4): 048101.

Krackhardt, D. 1999. The ties that torture: simmelian tie analysis in organizations. *Research in the Sociology of Organizations* 16 (1): 183–210.

Krivitsky, P. N., and M. S. Handcock. 2014. A separable model for dynamic networks. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 76 (1): 29–46.

Krivitsky, P. N., M. S. Handcock, A. E. Raftery, and P. D. Hoff. 2009. Representing degree distributions, clustering, and homophily in social networks with latent cluster random effects models. *Social networks* 31 (3): 204–213.

Latouche, P., E. Birmelé, and C. Ambroise. 2011. Overlapping stochastic block models with application to the french political blogosphere. *The Annals of Applied Statistics* 5:309–336.

Lusseau, D., K. Schneider, O. J. Boisseau, P. Haase, E. Slooten, and S. M. Dawson. 2003. The bottlenose dolphin community of doubtful sound features a large proportion of long-lasting associations. *Behavioral Ecology and Sociobiology* 54 (4): 396–405.

MacRae, D. 1960. Direct factor analysis of sociometric data. *Sociometry*:360–371.

Mariadassou, M., and C. Matias. 2015. Convergence of the groups posterior distribution in latent or stochastic block models. *Bernoulli* 21 (1): 537–573.

Meester, R. 1996. *Continuum percolation*. 119. Cambridge University Press.

Michael, J. H., and J. G. Massey. 1997. Modeling the communication network in a sawmill. *Forest Products Journal* 47 (9): 25–30.

Milgram, S. 1967. The small world problem. *Psychology today* 2 (1): 60–67.

Newman, M. E. J. 2002a. Assortative mixing in networks. *Physical review letters* 89 (20): 208701.

Newman, M. E. J. 2006. Finding community structure in networks using the eigenvectors of matrices. *Physical review E* 74 (3): 036104.

Newman, M. E. J. 2003a. Properties of highly clustered networks. *Physical Review E* 68 (2): 026121.

Newman, M. E. J. 2002b. Random graphs as models of networks. *arXiv:cond-mat/0202208*.

Newman, M. E. J. 2009. Random graphs with clustering. *Physical review letters* 103 (5): 058701.

Newman, M. E. J. 2003b. The structure and function of complex networks. *SIAM review* 45 (2): 167–256.

Newman, M. E. J. 2001. The structure of scientific collaboration networks. *Proceedings of the National Academy of Sciences* 98 (2): 404–409.

Newman, M. E. J., and J. Park. 2003. Why social networks are different from other types of networks. *Physical Review E* 68 (3): 036122.

Newman, M. E. J., S. H. Strogatz, and D. J. Watts. 2001. Random graphs with arbitrary degree distributions and their applications. *Physical Review E* 64 (2): 026118.

Nowicki, K., and T. A. B. Snijders. 2001. Estimation and prediction for stochastic block-structures. *Journal of the American Statistical Association* 96 (455): 1077–1087.

Olhede, S. C., and P. J. Wolfe. 2012. Degree-based network models. *arXiv:1211.6537*.

Padgett, J. F. 1994. Marriage and elite structure in reinassance florence; 1282-1500. *Redes: revista hispana para el análisis de redes sociales* 21:71–97.

Penrose, M. D. 1991. On a continuum percolation model. *Advances in applied probability* 23:536–556.

Perry, P. O., and P. J. Wolfe. 2013. Point process modelling for directed interaction networks. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 75 (5): 821–849.

Raftery, A. E., X. Niu, P. D. Hoff, and K. Y. Yeung. 2012. Fast inference for the latent space network model using a case-control approximate likelihood. *Journal of Computational and Graphical Statistics* 21 (4): 901–919.

Sampson, S. F. 1968. A novitiate in a period of change: an experimental and case study of social relationships. PhD thesis, Cornell University, September.

Schweinberger, M., and M. S. Handcock. 2015. Local dependence in random graph models: characterization, properties and statistical inference. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 77 (3): 647–676.

Shalizi, C. R., and A. Rinaldo. 2013. Consistency under sampling of exponential random graph models. *The Annals of Statistics* 41 (2): 508–535.

Söderberg, B. 2002. General formalism for inhomogeneous random graphs. *Physical review E* 66 (6): 066121.

Sweet, T. M., A. C. Thomas, and B. W. Junker. 2013. Hierarchical network models for education research: hierarchical latent space models. *Journal of Educational and Behavioral Statistics* 38:295–318.

Wang, H., M. Tang, Y. Park, and C. E. Priebe. 2014. *IEEE Transactions on Signal Processing* 62:703–717.

Watts, D. J., and S. H. Strogatz. 1998. Collective dynamics of small-world networks. *Nature* 393 (6684): 440–442.

Williams, R. J., and N. D. Martinez. 2000. Simple rules yield complex food webs. *Nature* 404 (6774): 180–183.