Network Science **4** (4): 407–432, 2016. © Cambridge University Press 2016
doi:10.1017/nws.2016.23

Properties of latent variable network models

RICCARDO RASTELLI and NIAL FRIEL
School of Mathematics and Statistics, University College Dublin, Dublin, Ireland
Insight: Centre for Data Analytics, Ireland
(e-mail: riccardo.rastelli@ucdconnect.ie; nial.friel@ucd.ie*)*

ADRIAN E. RAFTERY
Department of Statistics and Sociology, University of Washington, Seattle, USA
(e-mail: raftery@u.washington.edu*)*

Abstract

We derive properties of latent variable models for networks, a broad class of models that includes the widely used latent position models. We characterize several features of interest, with particular focus on the 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 model with random effects, and show that it can represent the heavy-tailed degree distributions, positive asymptotic clustering coefficients, and small-world behaviors that often occur in observed social networks. Finally, we illustrate the ability of the models to capture important features of real networks through several well-known datasets.

Keywords: *fitness models, latent position models, latent variable models, random graphs, social networks*

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 & Wolfe, 2013), technological networks such as the World Wide Web (Albert et al., 1999), and biological networks such as neural networks (Watts & Strogatz, 1998), food webs (Williams & Martinez, 2000), and protein–protein interaction networks (Raftery et al., 2012).

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

Networks are typically modeled 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 realization 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), or clustering (Newman, 2009) and small-world behavior (Watts & 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 behavior and small-world behavior to be assessed (Kiss & Green, 2008; Newman, 2009; Watts & Strogatz, 1998).

On the other hand, various statistical models have been proposed, including Exponential Random Graph Models (Frank & Strauss, 1986; Caimo & Friel, 2011; Krivitsky & Handcock, 2014), latent stochastic blockmodels (Nowicki & Snijders, 2001; Latouche et al., 2011; Airoldi et al., 2008), and latent position models (LPMs) (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 & Matias, 2012; Mariadassou & Matias, 2015). Moreover, recent developments (Chatterjee & Diaconis, 2013; Shalizi & Rinaldo, 2013; Schweinberger & 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 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 & Ward, 2014), trophic food webs (Chiu & Westveld, 2011; Chiu & Westveld, 2014), signal processing (Wang et al., 2014), and education research (Sweet et al., 2013).

Analytic expressions for the properties of this model in its original form are hard to derive. Therefore, we propose a new but closely related model, the Gaussian latent position model (Gaussian LPM). This yields tractable analytic expressions or asymptotic approximations for several important properties, including a complete characterization of the degree distribution, the clustering coefficient, and the distribution of path lengths. The availability of analytic expressions facilitates the analysis of large graphs, since the computational complexity required for simulations is greatly reduced.

One result is that the Gaussian LPM can capture transitivity in large networks, because its clustering coefficient is strictly positive for graphs of any size. This contrasts with the Erdős–Rényi Model whose clustering coefficient converges to zero when the number of nodes increases and the average degree is kept constant.

An implication of our results is that the LPM in its original form cannot represent heavy-tailed degree distributions, such as power-law behavior, or small-world behavior, as measured by the average path length (APL). Therefore, we introduce the Gaussian latent position model with random effects (GLPMRE), and show that it can overcome these limitations and capture important features of large-size real networks. These results suggest that the GLPMRE may be a good model for social networks.

https://doi.org/10.1017/nws.2016.23 Published online by Cambridge University Press

The main contributions of the paper are described in Sections 3 and 4, where a thorough analysis of Gaussian LPMs is given; as well as in Section 5, where the appealing properties of GLPMREs are shown through simulation studies and examples.

2 Latent variable network models

2.1 Notation and model assumptions

Here, we introduce our notation and define the various LVMs for networks that we consider.

We denote a binary graph by $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ is the set of node labels and $\mathcal{E}$ is the set of edges. We focus on undirected graphs with no self-edges such that

$$\forall (i,j) \in \mathcal{U} := \{(i,j) \in \mathcal{V} \times \mathcal{V} : 1 \leqslant i < j \leqslant n\} : y_{ij} = y_{ji} \tag{1}$$

$$\forall i \in \mathcal{V} : y_{ii} = 0 \tag{2}$$

We now propose a characterization of these network models using several modeling assumptions.

A1. *A latent variable $\mathbf{Z}_i \in \mathcal{Z}$ is associated to node i for every $i \in \mathcal{V}$, for a continuous or discrete set $\mathcal{Z}$. The set $\mathcal{M} = \{\mathbf{z}_1, \ldots, \mathbf{z}_n\}$ consists of realizations of independent and identically distributed latent random variables, where each $\mathbf{Z}$ is distributed according to the probability measure $p(\cdot)$.*

A2. *Edges are assumed to be conditionally independent given the latent variables. Hence, $\forall (i,j) \in \mathcal{U}$, Y_{ij} is a Bernoulli random variable such that*

$$Pr\left(Y_{ij} = 1 | \mathbf{z}_i, \mathbf{z}_j\right) = 1 - Pr\left(Y_{ij} = 0 | \mathbf{z}_i, \mathbf{z}_j\right) = r\left(\mathbf{z}_i, \mathbf{z}_j\right) \tag{3}$$

Definition 1
An LVM is a network model satisfying **A1** and **A2**.

Erdős–Rényi random graphs are a special case of LVMs where the connection probability function of Equation (3) is constant with respect to the latent information. The family of LVMs also includes the random connection models of Meester and Roy (1996), the fitness models of Caldarelli et al. (2002) and Söderberg (2002), the LPMs of Hoff et al. (2002), Handcock et al. (2007), and Krivitsky et al. (2009), and the stochastic blockmodels of Nowicki and Snijders (2001).

We now define our proposed Gaussian LPMs, which form another special case of the LVM:

A3. *The realized latent variables $\mathcal{M}$ in* **A1** *are points in the Euclidean space $\mathbb{R}^d$, for a fixed d, and they are normally distributed:*

$$p\left(\mathcal{M} | \gamma\right) = \prod_{i=1}^{n} f_d\left(\mathbf{z}_i; \mathbf{0}, \gamma\right) = \prod_{i=1}^{n} (2\pi\gamma)^{-\frac{d}{2}} \exp\left\{-\frac{1}{2\gamma}\mathbf{z}_i^t \mathbf{z}_i\right\} \tag{4}$$

In Equation (4), γ is a positive real parameter and $f_d(\cdot\,; \boldsymbol{\mu}, \gamma)$ is the multivariate Gaussian density function with mean vector $\boldsymbol{\mu}$ and covariance matrix $\gamma \mathbb{I}_d$, where $\mathbb{I}_d$ is the $d \times d$ identity matrix and $\mathbf{X}^t$ denotes the transpose of the matrix or vector $\mathbf{X}$.

https://doi.org/10.1017/nws.2016.23 Published online by Cambridge University Press

A4. *Given* $\varphi > 0$, $\tau \in [0,1]$*:*

$$r\left(\mathbf{z}_i, \mathbf{z}_j\right) = \tau \exp\left\{-\frac{\left(\mathbf{z}_i - \mathbf{z}_j\right)^t \left(\mathbf{z}_i - \mathbf{z}_j\right)}{2\varphi}\right\} \tag{5}$$

Definition 2
A Gaussian LPM is a network model satisfying **A1**–**A2**–**A3**–**A4**.

The Gaussian LPM differs from the original LPM of Hoff et al. (2002), in that the logistic connection function for the edges is replaced by a non-normalized Gaussian density. From now on, we will refer to the original LPM as the Logistic LPM.

2.1.1 Extensions of latent position models

Two extensions of the Logistic LPM were proposed by Handcock et al. (2007) and Krivitsky et al. (2009). In the former, clustering was introduced through a mixture distribution of the latent nodal positions, while in the latter nodal random effects were introduced to capture degree heterogeneity. In a similar fashion, we propose the following two extensions to the Gaussian LPM:

Definition 3
A Gaussian latent position cluster model (GLPCM) is a network model satisfying **A1**–**A2**–**A4** and such that the latent positions are distributed according to a finite mixture of Gaussian distributions, i.e.

$$p\left(\mathscr{M}|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\gamma}, G\right) = \prod_{i=1}^{n}\left[\sum_{g=1}^{G} \pi_g f_d\left(\mathbf{z}_i; \boldsymbol{\mu}_g, \gamma_g\right)\right] \tag{6}$$

where $\boldsymbol{\pi}$ are the mixture weights, $\boldsymbol{\mu}$ and $\boldsymbol{\gamma}$ are the parameters for the components, and G is the number of mixture components.

Definition 4
A GLPMRE is a network model satisfying **A1**–**A2**–**A3** such that

$$Pr\left(Y_{ij} = 1|\mathbf{z}_i, \mathbf{z}_j, \varphi_i, \varphi_j, \tau\right) = \tau \exp\left\{-\frac{1}{2\left(\varphi_i + \varphi_j\right)^2}(\mathbf{z}_i - \mathbf{z}_j)^t(\mathbf{z}_i - \mathbf{z}_j)\right\} \tag{7}$$

where, for every node $s \in \mathscr{V}$, $\varphi_s > 0$ is a random effect distributed according to an Inverse Gamma distribution with parameters β_0 and β_1:

$$p\left(\varphi_s|\beta_0, \beta_1\right) = \frac{\beta_1^{\beta_0}}{\Gamma\left(\beta_0\right)}\varphi^{-\beta_0 - 1} \exp\left\{-\frac{\beta_1}{\varphi_s}\right\} \tag{8}$$

2.2 Motivation for the Gaussian likelihood assumption

The Logistic LPM has been widely used in network models. Assumption **A4** introduces a new function to define the probability of edges, which is proportional to a Gaussian density. Although variations of the likelihood function have been proposed in the statistical community (Gollini & Murphy, 2016), the reasoning behind the Gaussian function mainly comes from the physics literature (Deprez &

https://doi.org/10.1017/nws.2016.23 Published online by Cambridge University Press

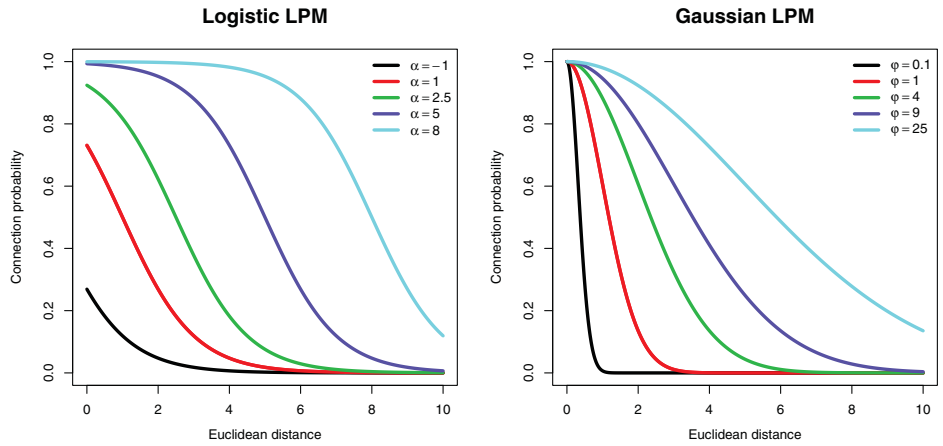

Fig. 1. Comparison between the Logistic and Gaussian connection functions, with $\tau = \gamma = 1$. As a function of the Euclidean distance between the nodes, in both cases, the likelihood of an edge reaches its maximum when the distance is null, and decreases to zero as the distance increases. (Color online)

Wüthrich, 2013; Penrose, 1991; Meester & Roy, 1996). The main advantage of using the Gaussian function is that it leads to tractable expressions for several theoretical properties of interest.

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

The parameter φ relates the probability of there being an edge to the distance between latent positions: The larger φ is, the more long range edges are supported. Note that the Erdős–Rényi random graph with connection probability τ is recovered as a special case for $\varphi \to \infty$.

Essentially, the main difference between the Gaussian and logistic likelihoods lies in the different behaviors as functions of the distance between nodes (Figure 1).

3 Theoretical results

3.1 Properties of the degrees

The degree of an arbitrary actor s is a discrete random variable defined by $D_s = \sum_{j \in \mathscr{V}} Y_{sj}$. We denote by $\mathbf{p} = (p_0, \ldots, p_{n-1})$ the degree distribution of a node chosen at random, i.e. $p_k = Pr(D = k)$, $\forall k = 0, \ldots, n-1$. To study the degree distribution of LVMs (and hence LPMs, as a special case), we propose a framework resembling that of Newman et al. (2001), which relies on Probability Generating Functions (PGFs).

The study will focus on the following quantities:

- **Q1:** $\theta(\mathbf{z}_s)$, defined as the probability that an actor chosen at random is a neighbor of a node with latent information $\mathbf{z}_s$.
- **Q2:** The PGF of the degree of a randomly chosen actor, $G(x) = \sum_{k=0}^{n-1} x^k p_k$.

https://doi.org/10.1017/nws.2016.23 Published online by Cambridge University Press

- **Q3:** The rth factorial moment, c_r, of the degree of a randomly chosen actor.[1]
- **Q4:** The expectation of the degree of a random node: $\bar{k}$, which is equal to the first factorial moment, c_1.
- **Q5:** The values of p_k, for every $k = 0, \ldots, n-1$.
- **Q6:** $\bar{k}(\mathbf{z}_s)$, defined as the expected degree of a node with latent information $\mathbf{z}_s$.
- **Q7:** $\bar{k}_{nn}(\mathbf{z}_s)$, defined as the expected degree of a random neighbor of a node with latent information $\mathbf{z}_s$.
- **Q8:** $\bar{k}_{nn}(k)$, defined as the expected degree of a random neighbor of a node with degree k.

The following main result characterizes these quantities for a LVM:

Theorem 1
For an LVM, the following results hold:

$$\textbf{Q1: } \theta(\mathbf{z}_s) = \int_{\mathscr{Z}} p\left(\mathbf{z}_j\right) r\left(\mathbf{z}_s, \mathbf{z}_j\right) d\mathbf{z}_j \tag{9}$$

$$\textbf{Q2: } G(x) = \int_{\mathscr{Z}} p\left(\mathbf{z}_s\right) \left[x\theta(\mathbf{z}_s) + 1 - \theta(\mathbf{z}_s)\right]^{n-1} d\mathbf{z}_s \tag{10}$$

$$\textbf{Q3: } c_r = \frac{\partial^r G}{\partial x^r}(1) = \frac{(n-1)!}{(n-r-1)!} \int_{\mathscr{Z}} p\left(\mathbf{z}_s\right) \theta(\mathbf{z}_s)^r d\mathbf{z}_s \tag{11}$$

$$\textbf{Q4: } \bar{k} = (n-1) \int_{\mathscr{Z}} p\left(\mathbf{z}_s\right) \theta(\mathbf{z}_s) d\mathbf{z}_s \tag{12}$$

$$\textbf{Q5: } p_k = \int_{\mathscr{Z}} p\left(\mathbf{z}_s\right) \binom{n-1}{k} \theta(\mathbf{z}_s)^k \left[1 - \theta(\mathbf{z}_s)\right]^{n-k-1} d\mathbf{z}_s \tag{13}$$

$$\textbf{Q6: } \bar{k}(\mathbf{z}_s) = (n-1)\theta(\mathbf{z}_s) \tag{14}$$

$$\textbf{Q7: } \bar{k}_{nn}(\mathbf{z}_s) = 1 + \frac{(n-2)}{\theta\left(\mathbf{z}_s\right)} \int_{\mathscr{Z}} p\left(\mathbf{z}_j\right) r\left(\mathbf{z}_s, \mathbf{z}_j\right) \theta\left(\mathbf{z}_j\right) d\mathbf{z}_j \tag{15}$$

$$\textbf{Q8: } \bar{k}_{nn}(k) = \frac{1}{p_k} \int_{\mathscr{Z}} p(\mathbf{z}_j) \binom{n-1}{k} \theta(\mathbf{z}_j)^k \left[1 - \theta(\mathbf{z}_j)\right]^{n-k-1} \bar{k}_{nn}(\mathbf{z}_j) d\mathbf{z}_j \tag{16}$$

The proof of Theorem 1 is given in the Supplementary Material.

Remark 1
Equation (16) is a generalization of a result of Boguná and Pastor-Satorras (2003), who introduced a general framework to study the degree correlations for the fitness model of Caldarelli et al. (2002) and Söderberg (2002).

Remark 2
Particular instances of some of the results of Theorem 1 have been shown in Olhede and Wolfe (n.d.) for fitness models and by Channarond et al. (2012) and Daudin et al. (2008) for stochastic block models, without resorting to PGFs. Theorem 1 encompasses those as special cases and extends the range of results shown.

We now apply these results to Gaussian LPMs. Proofs are shown in the Supplementary Material.

[1] The rth factorial moment of a discrete random variable D is defined as $\mathbb{E}\left[D\left(D-1\right)\cdots\left(D-r+1\right)\right]$.

https://doi.org/10.1017/nws.2016.23 Published online by Cambridge University Press

Corollary 1
In a Gaussian LPM, the following quantities have an explicit form:

$$\textbf{Q1: } \theta(\mathbf{z}_s) = \tau \left(\frac{\varphi}{\gamma + \varphi} \right)^{\frac{d}{2}} \exp \left\{ -\frac{1}{2(\gamma + \varphi)} \mathbf{z}_s^t \mathbf{z}_s \right\} \tag{17}$$

$$\textbf{Q3: } c_r = \frac{\partial^r G}{\partial x^r}(1) = \frac{(n-1)!}{(n-r-1)!} \tau^r \left\{ \frac{\varphi^r}{(\gamma + \varphi)^{r-1} [(r+1)\gamma + \varphi]} \right\}^{\frac{d}{2}} \tag{18}$$

$$\textbf{Q4: } \bar{k} = (n-1)\tau \left\{ \frac{\varphi}{2\gamma + \varphi} \right\}^{\frac{d}{2}} \tag{19}$$

$$\textbf{Q7: } \bar{k}_{nn}(\mathbf{z}_s) = 1 + \bar{k} \left(\frac{n-2}{n-1} \right) \frac{f_d \left(\mathbf{z}_s ; \mathbf{0}, \frac{\gamma^2 + 3\gamma\varphi + \varphi^2}{2\gamma + \varphi} \right)}{f_d (\mathbf{z}_s ; \mathbf{0}, \gamma + \varphi)} \tag{20}$$

Note that $\theta(\cdot)$ has an explicit expression, so evaluation of the quantities in **Q2**, **Q5**, and **Q8** boils down to an approximation of a single integral.

Corollary 2
In the GLPCM, the following results hold:

$$\textbf{Q1: } \theta(\mathbf{z}_s) = \tau (2\pi\varphi)^{\frac{d}{2}} \sum_{g=1}^{G} \pi_g f_d \left(\mathbf{z}_s ; \boldsymbol{\mu}_g, \gamma_g + \varphi \right) \tag{21}$$

$$\textbf{Q4: } \bar{k} = (n-1)\tau (2\pi\varphi)^{\frac{d}{2}} \sum_{g=1}^{G} \sum_{h=1}^{G} \pi_g \pi_h f_d \left(\mu_g - \mu_h ; \mathbf{0}, \gamma_g + \gamma_h + \varphi \right) \tag{22}$$

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

$$\textbf{Q5: } p_k = \int_{\mathbb{R}^d} \left[\sum_{g=1}^{G} \pi_g f_d \left(\mathbf{z}_s ; \mu_g, \gamma_g \right) \right] \binom{n-1}{k} \theta(\mathbf{z}_s)^k \left[1 - \theta(\mathbf{z}_s) \right]^{n-k-1} d\mathbf{z}_s \tag{23}$$

In the GLPMRE, none of the equations can be written explicitly, since the integrals over the random effects cannot be evaluated analytically. However, we will use numerical approximations to evaluate the intractable integrals appearing in **Q1** and **Q3**, thereby characterizing the properties of these models.

Remark 3
The advantage of using the Gaussian function rather than the Logistic function of Hoff et al. (2002) is highlighted in Corollary 1. Under the Gaussian hypothesis, several of the integrals of Equations (9)–(16) can be evaluated analytically since they become convolutions of two Gaussian densities, which can be evaluated analytically 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 with numerical approximations, since the number of integrals to evaluate is constant (i.e. it depends on d, but not on n).

Remark 4
In Gaussian LPMs, a non-identifiability issue arises between the parameters φ and γ, since the factorial moments depend only on their ratio, φ/γ. We keep both parameters, so 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 GLPCM and the GLPMRE.

https://doi.org/10.1017/nws.2016.23 Published online by Cambridge University Press

3.2 Clustering coefficient

We define the clustering coefficient as the probability that, if nodes i and j are connected, and nodes j and k are connected, then nodes i and k are connected. In the LVM, the clustering coefficient $\mathscr{C}$ can be written as

$$\mathscr{C} = \frac{\int_{\mathscr{Z}}\int_{\mathscr{Z}}\int_{\mathscr{Z}} p(\mathbf{z}_i)p(\mathbf{z}_k)p(\mathbf{z}_j)r(\mathbf{z}_i,\mathbf{z}_k)\, r\left(\mathbf{z}_k,\mathbf{z}_j\right) r\left(\mathbf{z}_j,\mathbf{z}_i\right) d\mathbf{z}_i d\mathbf{z}_k d\mathbf{z}_j}{\int_{\mathscr{Z}}\int_{\mathscr{Z}}\int_{\mathscr{Z}} p(\mathbf{z}_i)p(\mathbf{z}_k)p(\mathbf{z}_j)r(\mathbf{z}_i,\mathbf{z}_k)\, r\left(\mathbf{z}_k,\mathbf{z}_j\right) d\mathbf{z}_i d\mathbf{z}_k d\mathbf{z}_j} \tag{24}$$

Our definition is not the same as some others, such as the global clustering coefficient of Newman (2003a) or the local clustering coefficient of Watts and Strogatz (1998). However, our definition does represent one idea of clustering in a network, and it also allows theoretical evaluation. The following result characterizes the clustering coefficient for Gaussian LPMs.

Proposition 1

The clustering coefficient of a Gaussian LPM is

$$\mathscr{C} = \tau\left(\frac{\gamma+\varphi}{3\gamma+\varphi}\right)^{\frac{d}{2}} \tag{25}$$

A proof of Proposition 1 is provided in the Supplementary Material. Equation (25) gives an exact value of the clustering coefficient of Gaussian LPMs of any size. This contrasts with many other network models, for which the clustering coefficient can only be recovered asymptotically. Some interesting consequences of Equation (25) will be illustrated in Section 4.3.

3.3 Connectivity properties

We now characterize the connectivity structure of networks drawn from the Gaussian LPM. To do so, we first define the notion of a path for a random graph, and then give a general result about the connection of two nodes in Gaussian LPMs, once their latent positions are known.

Definition 5 (*Path*)

In a network model, 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. $y_{i_0 i_1} = y_{i_1 i_2} = \cdots = y_{i_{k-1} i_k} = 1$.

In an LVM, 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

$$\xi_k(\mathbf{z}_i,\mathbf{z}_j) = \int_{\mathscr{Z}}\cdots\int_{\mathscr{Z}} p(\mathbf{z}_1)\ldots p(\mathbf{z}_{k-1})r(\mathbf{z}_i,\mathbf{z}_1)\, r(\mathbf{z}_1,\mathbf{z}_2)\cdots r\left(\mathbf{z}_{k-1},\mathbf{z}_j\right) d\mathbf{z}_1\cdots d\mathbf{z}_{k-1} \tag{26}$$

For a Gaussian LPM, the integrals on the right-hand side of Equation (26) involve Gaussian kernels only, and therefore can be evaluated exactly. The following proposition gives a more explicit expression.

https://doi.org/10.1017/nws.2016.23 Published online by Cambridge University Press

Proposition 2
In a Gaussian LPM, let $\mathbf{z}_i \in \mathbb{R}^d$, $\mathbf{z}_j \in \mathbb{R}^d$, and $\xi_k(\mathbf{z}_i, \mathbf{z}_j)$ be defined as in Equation (26), for any $k = 1, 2, \ldots, n-1$,. Define the following recurrence relations:

$$\begin{cases} h_{r+1} &= h_r \alpha_r^{-d} \tau (2\pi\varphi)^{\frac{d}{2}} f_d\left(\mathbf{z}_i; \mathbf{0}, \frac{\omega_r+\gamma}{\alpha_r^2}\right) \\ \alpha_{r+1} &= \frac{\alpha_r \gamma}{\omega_r+\gamma} \\ \omega_{r+1} &= \frac{\omega_r\varphi+\omega_r\gamma+\gamma\varphi}{\omega_r+\gamma} \end{cases}, \text{ where } \begin{cases} h_1 &= \tau (2\pi\varphi)^{\frac{d}{2}} \\ \alpha_1 &= 1 \\ \omega_1 &= \varphi \end{cases} \tag{27}$$

Then,

$$\xi_k(\mathbf{z}_i, \mathbf{z}_j) = h_k f_d\left(\mathbf{z}_j - \alpha_k \mathbf{z}_i; \mathbf{0}, \omega_k\right), \quad \text{for } k = 1, 2, \ldots, n-1 \tag{28}$$

The proof of Proposition 2 is in the Supplementary Material. Proposition 2 is useful for studying the statistical properties of path lengths for Gaussian LPMs, which we develop in Section 4.4.

4 Properties of realized networks

A drawback of all LPMs is that, given the complete set of latent positions, the evaluation of the likelihood for the corresponding realized 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 thousand nodes. This impasse is addressed in Raftery et al. (2012), where a computational approximation is proposed to overcome this difficulty. The computational issue extends also to the simulation of LPMs, which is usually performed in two sequential steps: First, 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 simulation-based study of the properties of LPMs rather inefficient and limited to small graphs, only.

By contrast, the results presented in Theorem 1 and Corollaries involve either exact formulae, which have negligible computational cost, or integral approximations whose computational cost is independent of n. Hence, the analysis that we propose does not require any intensive calculation and can be performed on networks of any size.

4.1 Characterization of the degree distribution for the Gaussian LPM

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 behavior 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, the scale-free property is a feature that is prioritized in the design of network models (Barabási & 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, and can be approximated efficiently for any network size. Figure 2 shows approximate degree distributions for various choices of model parameters.

https://doi.org/10.1017/nws.2016.23 Published online by Cambridge University Press

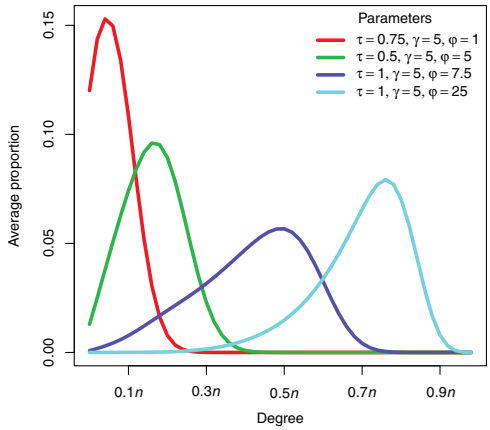

Fig. 2. Gaussian LPM: Approximate degree distribution for different sets of model parameters τ, γ, φ. (Color online)

The shapes of the theoretical degree distribution of Gaussian LPMs shown in Figure 2 suggest that left-skewed distributions are exhibited in denser networks. This may not be a desirable feature, since it suggests that Gaussian LPMs would not capture heavy tails and scale-free behaviors. Such left-skewed shapes do sometimes arise in social networks; however, 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). Dunbar (1992) has argued that there is a theoretical cognitive limit on the number of stable relationships that social actors can maintain. Hence, both power-law and non-power-law behaviors are of interest in statistical modeling of networks.

We now propose a more rigorous analysis of the degree distribution using the dispersion[2] and skewness indices[3], which can be evaluated through the exact formulae for the factorial moments in Equation (18).

Corollary 3

In a Gaussian LPM, the dispersion index is given by

$$\mathscr{D} = 1 + (n-2)\tau \left(\frac{\varphi(2\gamma+\varphi)}{(\gamma+\varphi)(3\gamma+\varphi)} \right)^{\frac{d}{2}} - (n-1)\tau \left(\frac{\varphi}{2\gamma+\varphi} \right)^{\frac{d}{2}} \tag{29}$$

The proof is given in the Supplementary Material.

[2] The dispersion index of a discrete random variable is equal to its variance divided by its mean. 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; hence, it qualifies as underdispersed.

[3] The skewness index of a random variable is equal to the third central moment divided by the cube of the standard deviation. In the case of degree distributions for networks, a positive value of the skewness index corresponds to shapes exhibiting a right tail heavier than the left one.

https://doi.org/10.1017/nws.2016.23 Published online by Cambridge University Press

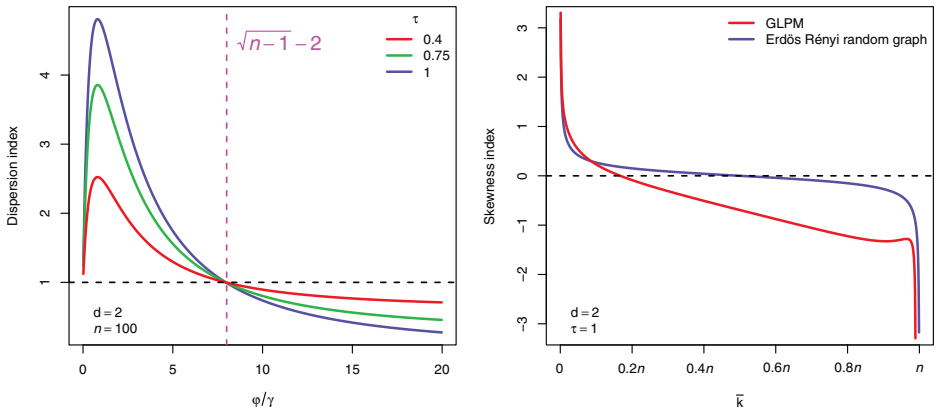

Fig. 3. Gaussian LPM: Left: Dispersion index versus the ratio between φ and γ. The vertical line is the threshold corresponding to a Poisson dispersion. For larger values of φ, the distributions arising are not more dispersed than an Erdős–Rényi random graph, asymptotically degenerating to this model as $\varphi \to \infty$. 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. (Color online)

Remark 5
The calculation of the skewness does not involve any simplification, and so it is omitted here.

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 summarized as follows:

- When $\varphi = \gamma(\sqrt{n-1}-2)$, the distribution has dispersion index 1, as a Poisson distributed variable.
- When $\varphi < \gamma(\sqrt{n-1}-2)$, the distribution has dispersion index greater than 1, yielding an overdispersed distribution.
- 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 characterization does not depend on τ.

The left panel of Figure 3 shows the dispersion as a function of the model parameters. 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, and below the threshold, then the model is not degenerate 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 behavior.

We now analyze the skewness index, which is useful for identifying asymmetries in overdispersed distributions. We expect a scale-free network to have a positive and relatively large skewness index, but Gaussian LPMs do not produce this behavior, as shown in the right panel of Figure 3. In Erdős–Rényi random graphs the degrees exhibit approximately a Poisson distribution, so p_k goes to zero at the rate $1/k!$,

https://doi.org/10.1017/nws.2016.23 Published online by Cambridge University Press

as n remains fixed and k tends to n. Thus, power-laws are not represented. The right panel in Figure 3 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 behavior. 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 non-trivially dispersed distribution between 0 and a maximum degree value, confirming the shapes already shown in Figure 2.

4.2 Degree correlations

In the study of networks, one is often interested in the mixing properties of the graph, where the mixing structure usually refers to the fact that nodes sharing common features are more likely to be linked. In the context of social networks, this behavior is called homophily.

A special case is mixing according to nodes' degrees, often called degree correlation. For example, one may be interested in whether the degrees of two random nearest neighbors are positively or negatively correlated. Positive correlation, or assortative mixing of the degrees, is a recurring feature in social networks (Newman & Park, 2003; Newman, 2002), in contrast to many other kinds of networks (World Wide Web, protein interactions, food webs; see Newman (2003b)), which typically exhibit 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 (20) shows that the Average Nearest Neighbor Degree (ANND) of an arbitrary node i can be written explicitly as a function of its latent position $\mathbf{z}_i$. The left panel of Figure 4 displays this function in terms of the distance between $\mathbf{z}_i$ and the center of the latent space.

It is not surprising that nodes located closer to the center have highly connected neighbors. Instead, Equation (16) provides a less explicit formula for the ANND as a function of the degree of node i, rather than its distance from the center. This quantity can be efficiently approximated for every degree value. The right panel of Figure 4 represents this case. The average degree of the neighbors of a node of degree k, $\bar{k}_{nn}(k)$, appears to be a non-decreasing 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 realized Gaussian LPM networks exhibit assortative mixing of the degrees, suggesting them to be well suited for social networks from this point of view (Newman & Park, 2003).

4.3 Asymptotics for the clustering coefficient

LPMs capture transitivity in a natural way. When two actors have a neighbor in common, it is to be expected that the three corresponding nodes will be close in the latent space, making triangles more likely. We will now show how Proposition 1 provides a more precise basis for this intuition.

https://doi.org/10.1017/nws.2016.23 Published online by Cambridge University Press

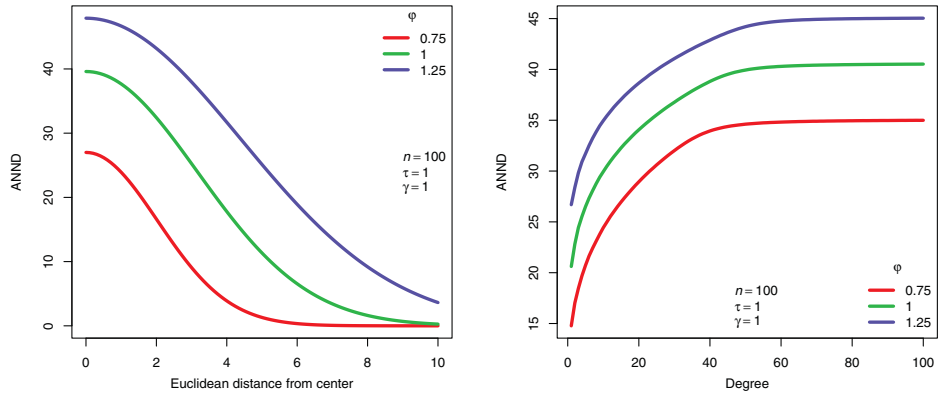

Fig. 4. Gaussian LPM: Left: Average degree of the nearest neighbors of a node as a function of its distance from the center. Nodes located in the center will more likely connect to high degree nodes. Right: Average degree of the closest neighbors as a function of the degree of a node. The ANND is clearly a non-decreasing function, verifying that Gaussian LPMs exhibit assortative mixing in the degrees of the nodes. (Color online)

One 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 realized network. We focus on the 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 $\mathscr{C} \to 0$ since $\mathscr{C} = p$. Hence, asymptotically, the clustering coefficient for Erdős–Rényi random graphs is zero.

In contrast, Gaussian LPMs can represent transitivity, even asymptotically. To see this, we again consider the situation where the average degree is kept fixed and n increases. In Gaussian LPMs, the average degree can be kept fixed in a number of ways as n increases. We choose to use the result in Equation (19) which allows us to have the same average degree $\bar{k}_0$ for every n, by imposing the following constraint:

$$\varphi = \frac{2\bar{k}_0^{\frac{2}{d}}\gamma}{(n-1)^{\frac{2}{d}}\tau^{\frac{2}{d}} - \bar{k}_0^{\frac{2}{d}}} \tag{30}$$

As n tends to infinity, the corresponding clustering coefficient converges to

$$\mathscr{C} = \frac{\tau}{3^{\frac{d}{2}}} \tag{31}$$

Hence, the limiting clustering coefficient has a non-zero value that can be as large as $3^{-\frac{d}{2}}$, suggesting that Gaussian LPMs are able to capture a persistent transitivity for networks of any size.

The asymptotically 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. Newman reviewed technical difficulties in the analysis of highly clustered models.

https://doi.org/10.1017/nws.2016.23 Published online by Cambridge University Press

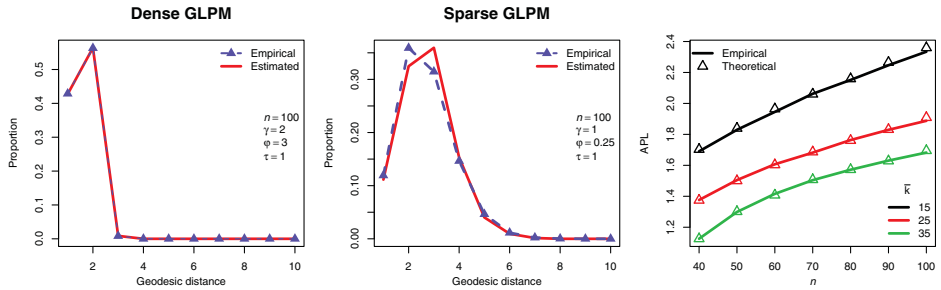

Fig. 5. Geodesic distances and average path lengths for the Gaussian LPM model. Left and center: Comparison between the theoretical values and average values obtained through simulations for the distribution of geodesic distances. The simulated networks are composed of 100 nodes. On the left panel, the graph is more dense (average degree is approximately 42) while the one in the center is sparser (average degree is approximately 11). Right: Comparison between average values of simulated networks (lines) and theoretical (triangles) values for the APL. The parameters τ and γ are set to 1. (Color online)

4.4 Path lengths

In this section, we study the distribution of shortest path lengths (geodesic distances) and characterize the behavior of the APL for Gaussian LPMs. We refer to the concept of small-world networks originally introduced in Watts and Strogatz (1998). As a comparison tool, we use again the Erdős–Rényi model, for which the APL is typically proportional to the log of the size of the network.

We denote by $\ell_k\left(\mathbf{z}_i, \mathbf{z}_j\right)$ the probability that k is the length of the shortest path between nodes i and j, located in $\mathbf{z}_i$ and $\mathbf{z}_j$, respectively, given that there exists at least one path connecting them.

A characterization of the geodesic distance for fitness models (hence including Erdős–Rényi models) appeared in Fronczak et al. (2004). Here, we follow the same reasoning, extending the method to Gaussian LPMs. As a result, we argue that for dense Gaussian LPM networks, the distribution of the geodesic distances is characterized by

$$\ell_k\left(\mathbf{z}_i, \mathbf{z}_j\right) \approx \exp\left\{-n^{k-1}\xi_{k-1}(\mathbf{z}_i, \mathbf{z}_j)\right\} - \exp\left\{-n^{k}\xi_{k}(\mathbf{z}_i, \mathbf{z}_j)\right\} \tag{32}$$

A demonstration of the procedure used to obtain Equation (32) is shown in the Supplementary Material.

In Figure 5, a comparison between the theoretical values and those obtained through simulations is shown. The first two panels of Figure 5 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.

Furthermore, once $\ell_k\left(\mathbf{z}_i, \mathbf{z}_j\right)$ 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 approximation with the results from a simulation study is shown in the right panel of Figure 5. As expected, the approximation is more accurate for graphs with a higher average degree. However, the results show that such an index

https://doi.org/10.1017/nws.2016.23 Published online by Cambridge University Press

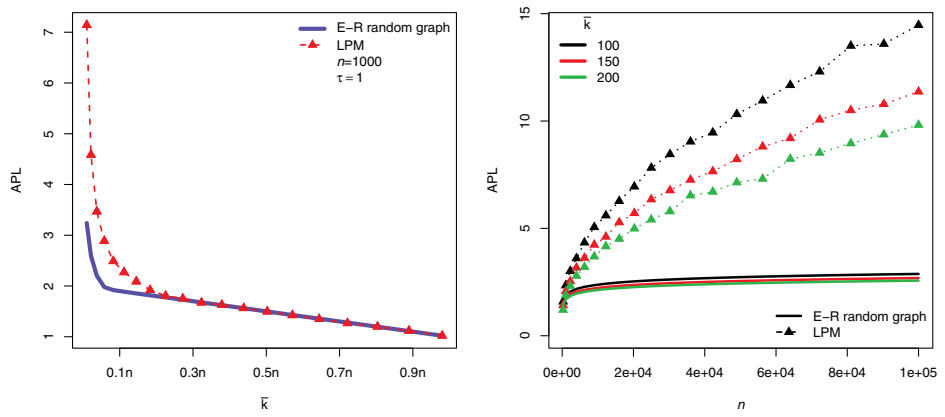

Fig. 6. Left: APL against the average degree of a 1,000-node network, compared with the corresponding Erdős–Rényi random graph. The two behaviors diverge for sparse graphs, in which case Gaussian LPMs exhibit a larger APL. Right: Asymptotic behavior for the APL is shown. Average degree of the network 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 the Gaussian LPM, and grows proportionally to a function which dominates the logarithm. (Color online)

is more tolerant when assumptions tend to be violated, possibly because the bias is limited when values are averaged.

Figure 6 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, Gaussian LPMs tend to behave more and more similarly to Erdős–Rényi random graphs. In the right panel of Figure 6, APL values are shown for larger Gaussian LPMs networks. In this case, the average degree is kept constant, highlighting the asymptotic behavior of the statistic.

APL values for the corresponding Erdős–Rényi random graphs are also shown in Figure 6. The Gaussian LPM networks typically have a higher APL, which grows faster than the logarithm of the size of the network.

Figure 7 illustrates a possible reason for this behavior. The Euclidean distance from a node to the center 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 behavior of Erdős–Rényi random graphs. Evidently, 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 GLPMRE, the connectivity parameter φ becomes node dependent, and is a realization of an Inverse Gamma distribution with parameters β_0 and β_1. An increase in φ mainly affects how prone the corresponding actor is to creating

https://doi.org/10.1017/nws.2016.23 Published online by Cambridge University Press

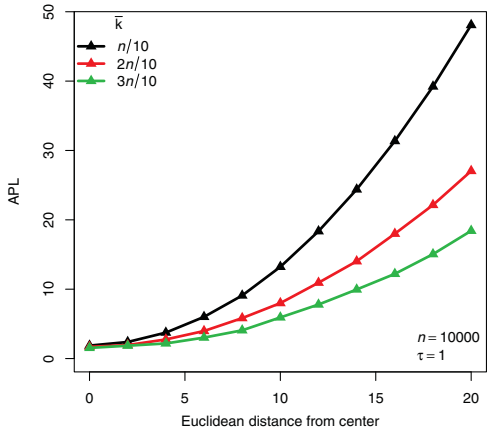

Fig. 7. Average geodesic distance from a node as a function of its distance from the center of the latent space. The network is composed of 10,000 nodes, with $\tau = 1$. Clearly, nodes which are closer to the center will be better positioned to reach easily many other nodes, thus having a smaller APL index. Such heterogeneity in the connectivity structure characterizes Gaussian LPMs and separates them from Erdős–Rényi random graphs, justifying the larger values for global APL. (Color online)

long-range connections, rather than short-range ones. This behavior 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 APL (Watts & Strogatz, 1998).

The results in **Q1** and **Q3** of Theorem 1 can be adapted to GLPMREs using numerical approximations: The latent positions can be integrated out analytically, and the random effects can then be integrated out numerically. Through such approximated quantities, the factorial moments of the degree of a random node can be characterized 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 in this framework τ does not play a crucial role, we fix it to 1.

Table 1 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 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 8 shows that an increase in the variance of the random effects does yield 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 9 shows the degree frequencies obtained through simulations of GLPMREs. The two panels on the left side of Figure 9 show the degree distribution for a Gaussian LPM

https://doi.org/10.1017/nws.2016.23 Published online by Cambridge University Press

Table 1. *Average degree of a network of* 100 *actors for different values of mean and variance of the nodal random effects. The variance does not have much impact on the average degree of the network. This suggests that any increase of skewness is not due to the network getting sparser.*

Mean \ Variance	0.0001	0.1	1	10	100	1,000	100,000
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

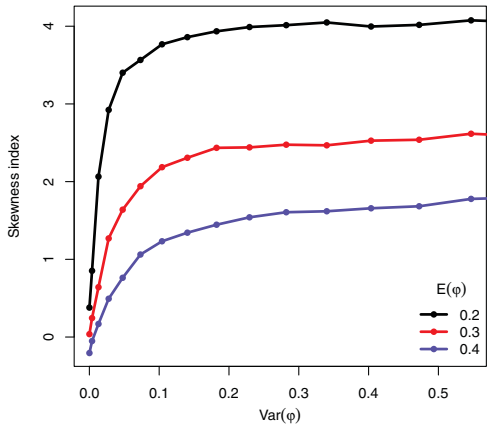

Fig. 8. Skewness index versus variance of nodal random effects. An increase in the variance of the random effects leads to an increase of the skewness index, corresponding to heavier tails. (Color online)

(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 indices 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 confirm that random effects can extend the family of networks represented using Gaussian LPMs. However, other features of interest are non-trivially influenced. Hence, we propose a simulation study to explore how random effects affect the asymptotic behavior of LPMRE with respect to small-world behavior and transitivity. Simulations of GLPMREs are inefficient, so the results are somewhat limited. However, such a procedure is the only feasible one, since theoretical results on the GLPMRE are not available.

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 behavior of APL and $\mathscr{C}$, we have held the average degree approximately constant by imposing $\gamma \propto n$, with n increasing.

https://doi.org/10.1017/nws.2016.23 Published online by Cambridge University Press

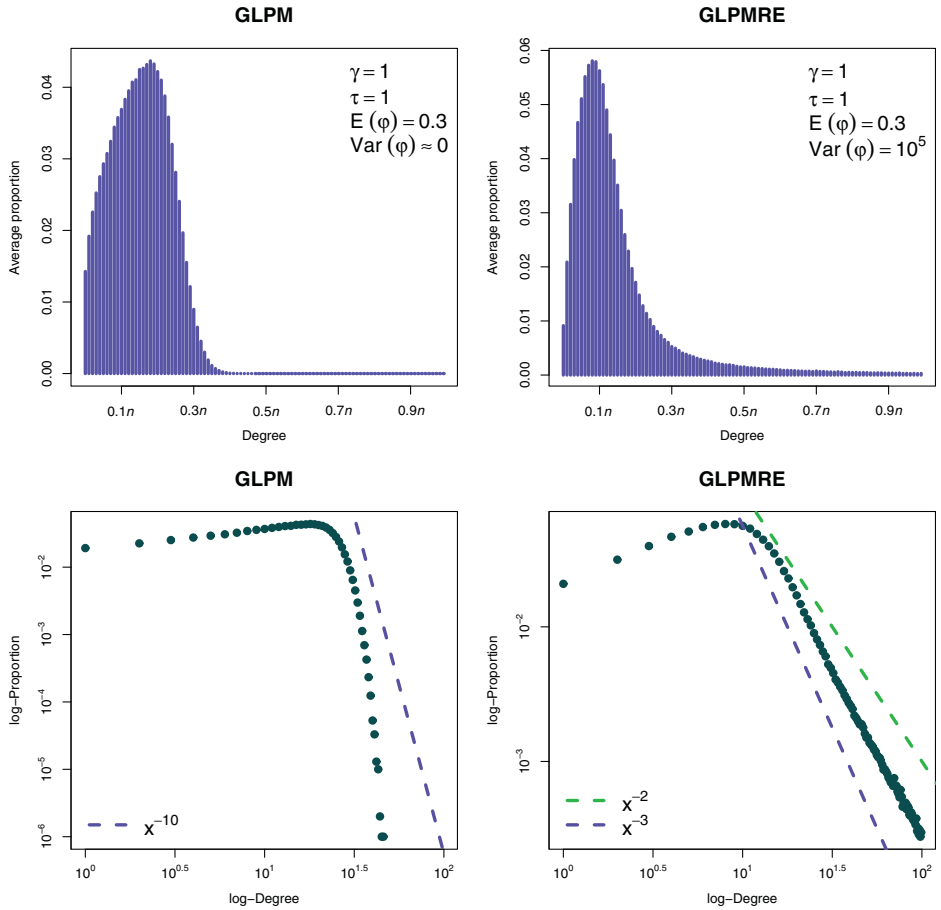

Fig. 9. Top: degree distributions for GLPMREs with null-variance random effects (left) and large-variance random effects (right). Bottom: corresponding degree distribution on the log–log scale. An increase in the variance of the random effects results in a heavier power-law tailed degree distribution. The average degrees are: $0.151n$ and $0.144n$ for the case on left and right, respectively, while skewness indices are -0.07 and 2.53, respectively. (Color online)

Figure 10 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, suggesting small-world behavior.

The right panel illustrates the asymptotic clustering coefficient estimated through simulations, showing that $\mathscr{C}$ tends to stabilize to a non-zero limiting value, which clearly depends on the variance of the random effects. Such an interaction between the presence of hubs and the clustering coefficient is not unexpected, since for an extreme case, the n-nodes star, $\mathscr{C}$ is equal to zero.

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

https://doi.org/10.1017/nws.2016.23 Published online by Cambridge University Press

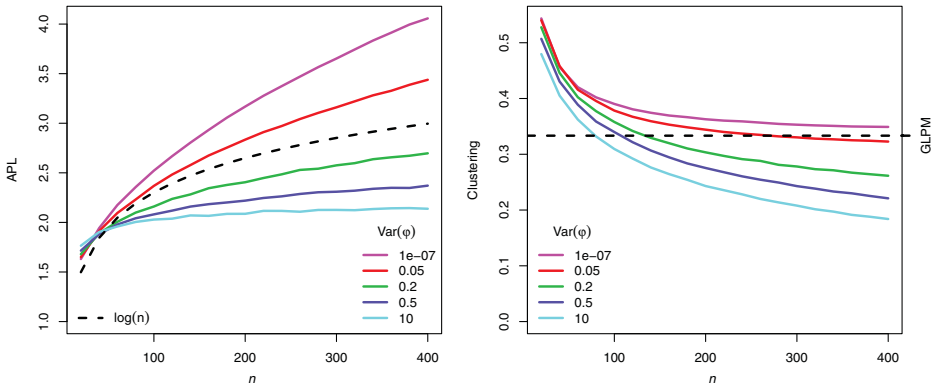

Fig. 10. APL (left) and clustering coefficient (right) as a function of n, holding an approximately constant average degree. The remaining model parameters are $\tau = 1$, $\mathbb{E}[\varphi] = 0.6$, and $\gamma = 0.05(n-1)$. The number of networks generated for each value of n is 1,000. The dashed black lines represent the log function and the asymptotic value for $\mathscr{C}$ under the Gaussian LPM for the left and right panel, respectively. (Color online)

6 Real data examples

We show in this section that several well-known real social networks have features that can be captured by a Gaussian LPM. We focus on 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 the Renaissance (Padgett & Ansell, 1993).
- **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**: Thirty-eight 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**: Thirty-six employees of a sawmill were asked to quantify the time they spent discussing work matters with each of their colleagues (Michael & 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 neighborhood called San Juan Sur (de Nooy et al., 2011).
- **Network sciences**: Coauthorship network of 1,589 scientists working on network theory and experiment (Newman, 2006).

https://doi.org/10.1017/nws.2016.23 Published online by Cambridge University Press

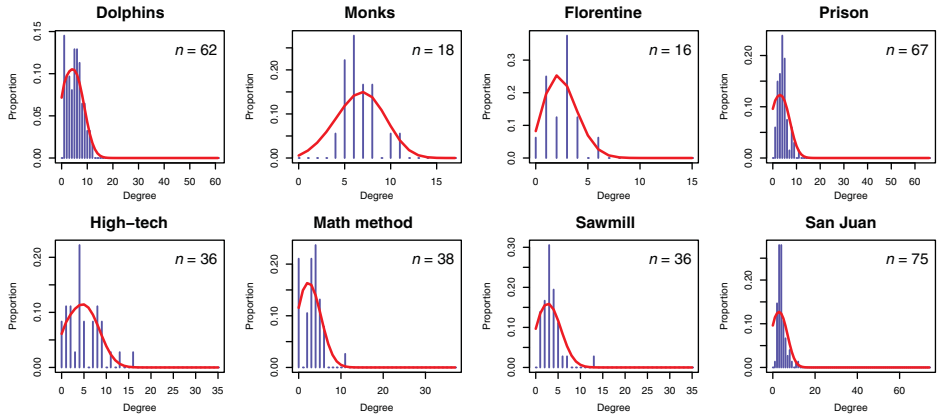

Fig. 11. Comparison between the observed degree distributions (blue bars) and the theoretical ones (red lines) for several small-size real social networks. Datasets used (from top left by row): Dolphins, Monks, Florentine, Prison, High-tech, Math method, Sawmill, San Juan. (Color online)

- **Geometry**: Coauthorship network of 7,343 scientists working on computational geometry (Batagelj & Mrvar, 2006).
- **Condensed Matter**: Coauthorships between 16,726 scientists posting preprints on the Condensed Matter E-Print Archive (Newman, 2001).
- **High energy**: Coauthorships between 27,770 scientists posting preprints on the High-Energy Theory E-Print Archive (Newman, 2001).

Some of the mentioned datasets are not binary and undirected. These have been transformed in order to conform to this requirement.

We have shown that a number of network statistics can be expressed through analytical formulae for the Gaussian LPM. These include the average degree $\bar{k}$, the clustering coefficient $\mathscr{C}$, the APL and the skewness index S. Here, we use an ad-hoc method to calibrate the parameters of a Gaussian LPM in the first eight datasets: Using numerical methods, the values of τ and φ that give the best match between the observed and theoretical $\bar{k}$ and $\mathscr{C}$ are obtained. In all of the cases, the matching is perfect (non-uniqueness of the solution is not relevant in this context). The parameter γ is fixed to 1. Then, for both the skewness and the APL indices, observed values are compared to the theoretical values corresponding to the optimal τ and φ, and to an interquantile range obtained from 1,000 networks simulated using the same parameters. Table 2 shows the results of this experiment.

It appears that the there is a good agreement between observed and theoretical values, with the exception of the skewness index, which is too large to be captured by the model in some of the networks. The observed and theoretical degree distributions of the networks are also in good agreement, as shown in Figure 11.

A similar approach was used to calibrate the GLPMRE model to the four larger datasets, to assess to what extent this model can capture the asymptotic scale-free decay of the degree distribution. The datasets are collaboration networks where nodes correspond to authors and two nodes are linked if the corresponding scientists published a paper as coauthors. All of these networks exhibit a power-law degree distribution, with different power orders, which can vary in the range 1 to

https://doi.org/10.1017/nws.2016.23 Published online by Cambridge University Press

Table 2. *Statistics for small-sized social networks. The average degree and the clustering coefficient are matched exactly in every case, while the observed skewness index and average path length are fairly close to the theoretical counterparts. The observed indices are compared with the theoretical values, and with the interval given by the* 0.05 *and* 0.95 *quantiles.*

Dolphins: $n = 62,\ \tau = 0.810,\ \varphi/\gamma = 0.232,\ \bar{k} = 5.13,\ \mathscr{C} = 0.31$				
	Obs.	**Th.**	**0.05 q**	**0.95 q**
Skewness	0.292	0.461	−0.105	0.656
APL	3.357	3.282	2.65	3.663
Monks: $n = 18,\ \tau = 0.763,\ \varphi/\gamma = 2.115,\ \bar{k} = 6.67,\ \mathscr{C} = 0.47$				
	Obs.	**Th.**	**0.05 q**	**0.95 q**
Skewness	0.877	−0.05	−0.845	0.606
APL	1.68	1.724	1.51	1.922
Flomarriage: $n = 16,\ \tau = 0.302,\ \varphi/\gamma = 2.460,\ \bar{k} = 2.5,\ \mathscr{C} = 0.19$				
	Obs.	**Th.**	**0.05 q**	**0.95 q**
Skewness	0.424	0.503	−0.458	1.097
APL	2.486	2.827	1.956	3.2
Prison: $n = 67,\ \tau = 0.776,\ \varphi/\gamma = 0.180,\ \bar{k} = 4.24,\ \mathscr{C} = 0.29$				
	Obs.	**Th.**	**0.05 q**	**0.95 q**
Skewness	0.855	0.562	−0.004	0.747
APL	3.355	3.831	2.916	4.311
High tech: $n = 36,\ \tau = 0.913,\ \varphi/\gamma = 0.376,\ \bar{k} = 5.06,\ \mathscr{C} = 0.37$				
	Obs.	**Th.**	**0.05 q**	**0.95 q**
Skewness	0.785	0.376	−0.306	0.615
APL	2.360	2.749	2.169	3.189
Math method: $n = 38,\ \tau = 0.616,\ \varphi/\gamma = 0.328,\ \bar{k} = 3.21,\ \mathscr{C} = 0.25$				
	Obs.	**Th.**	**0.05 q**	**0.95 q**
Skewness	0.654	0.612	−0.064	0.927
APL	2.644	3.480	2.474	4.045
Sawmill: $n = 36,\ \tau = 0.550,\ \varphi/\gamma = 0.436,\ \bar{k} = 3.44,\ \mathscr{C} = 0.23$				
	Obs.	**Th.**	**0.05 q**	**0.95 q**
Skewness	2.290	0.558	−0.119	0.919
APL	3.138	3.210	2.382	3.678
San Juan: $n = 75,\ \tau = 0.657,\ \varphi/\gamma = 0.186,\ \bar{k} = 4.13,\ \mathscr{C} = 0.25$				
	Obs.	**Th.**	**0.05 q**	**0.95 q**
Skewness	1.622	0.579	0.066	0.816
APL	3.485	3.883	3.034	4.326

https://doi.org/10.1017/nws.2016.23 Published online by Cambridge University Press

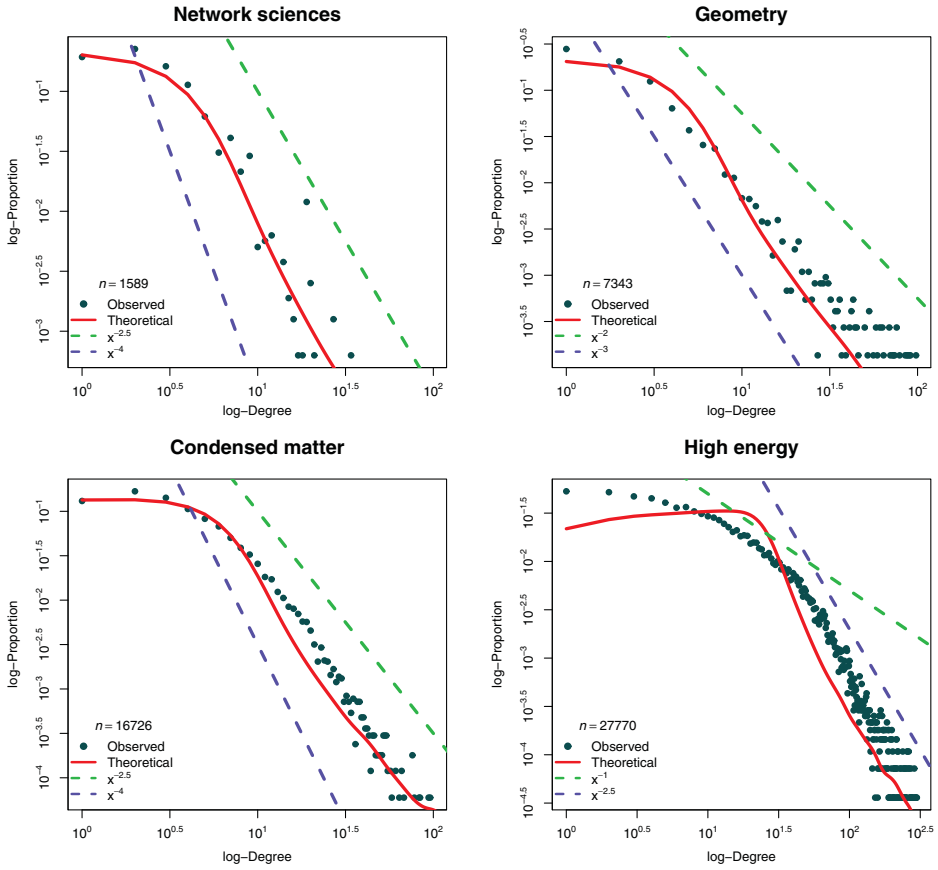

Fig. 12. Empirical (dots) and theoretical (red line) degree distributions on the log–log scale for four large citation networks. The datasets exhibit different asymptotic power-law orders. In all of the cases, GLPMREs appear to capture the scale-free behavior sufficiently well. Datasets used: Network sciences (top left), Geometry (top right), Condensed matter (bottom left), High energy (bottom right). (Color online)

4. Figure 12 shows the theoretical and observed degree distributions on the log–log scale, indicating that the asymptotic behavior is reasonably well captured by the model in all of the cases.

7 Conclusions

We have derived several properties of LPMs for networks, including their degree distribution, degree correlations, clustering coefficient, and the distribution of path lengths. We have also introduced a new class of LPMs, the Gaussian LPMs, which are more tractable analytically than other LPMs.

We have shown that Gaussian LPMs have an asymptotically strictly positive clustering coefficient, in contrast to other well-known models, such as Erdős–Rényi, whose clustering coefficient is asymptotically zero. This result suggests that Gaussian LPMs can generate highly clustered networks and that they can capture the persistent clustering behavior of large social networks.

https://doi.org/10.1017/nws.2016.23 Published online by Cambridge University Press

We have characterized the average degree of the nearest neighbors to a node, showing that positive degree correlations arise in Gaussian LPM networks. This is in line with observed social networks, where assortative mixing in the nodal degrees does occur.

We have also shown how the distribution of geodesic distances can be efficiently approximated, yielding an analysis of the asymptotic behavior of the APL.

In their basic form, Gaussian LPMs are not appropriate for modeling scale-free networks, since their degree distribution has a left-skewed and truncated shape. However, if they are modified by introducing node-specific random effects, yielding the GLPMRE model, we have shown that they can represent power-law distributions of different shapes in both simulated and real networks.

Although this work deals only with undirected graphs, the same results can be extended in a similar fashion to directed graphs.

Acknowledgments

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 Foundation Ireland grant 11/W.1/I2079 and by National Institutes of Health grant U54-HL127624. We are grateful to the Editor and two anonymous reviewers for very helpful comments.

Supplementary material

To view supplementary material for this article, please visit http://dx.doi.org/10.1017/nws.2016.23

References

Airoldi, E. M., Blei, D. M., Fienberg, S. E., & Xing, E. P. (2008). Mixed membership stochastic blockmodels. *Journal of Machine Learning Research*, **9**, 1981–2014.

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

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

Amaral, L. A. N., Scala, A., Barthelemy, M., & Stanley, H. E. (2000). Classes of small-world networks. *Proceedings of the National Academy of Sciences*, **97**(21), 11149–11152.

Ambroise, C., & Matias, C. (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., & Albert, R. (1999). Emergence of scaling in random networks. *Science*, **286**(5439), 509–512.

Batagelj, V., & Mrvar, A. (2006). Pajek datasets. http://vlado.fmf.uni-lj.si/pub/networks/data/

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

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

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

Cao, X., & Ward, M. D. (2014). Do democracies attract portfolio investment? Transnational portfolio investments modeled as dynamic network. *International Interactions*, **40**(2), 216–245.

Carlson, R. O. (1965). *Adoption of educational innovations*. Eugene, OR: Center for the Advanced Study of Educational Administration, University of Oregon.

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

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

Chiu, G. S., & Westveld, A. H. (2011). A unifying approach for food webs, phylogeny, social networks, and statistics. *Proceedings of the National Academy of Sciences*, **108**(38), 15881–15886.

Chiu, G. S., & Westveld, A. H. (2014). A statistical social network model for consumption data in trophic food webs. *Statistical Methodology*, **17**(4432), 139–160.

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

de Nooy, W., Mrvar, A., & Batgelj, V. (2011). *Exploratory social network analysis with Pajek* (2nd ed.). Cambridge, UK: Cambridge University Press.

Deprez, P., & Wüthrich, M. V. (2013). Scale-free percolation in continuum space. *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., & Strauss, D. (1986). Markov graphs. *Journal of the American Statistical Association*, **81**(395), 832–842.

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

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

Handcock, M. S., Raftery, A. E., & Tantrum, J. M. (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., Raftery, A. E., & Handcock, M. S. (2002). Latent space approaches to social network analysis. *Journal of the American Statistical Association*, **97**(460), 1090–1098.

Kiss, I. Z., & Green, D. M. (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., & Handcock, M. S. (2014). A separable model for dynamic networks. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **76**(1), 29–46.

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

Latouche, P., Birmelé, E., & Ambroise, C. (2011). Overlapping stochastic block models with application to the french political blogosphere. *Annals of Applied Statistics*, **5**(1), 309–336.

Lusseau, D., Schneider, K., Boisseau, O. J., Haase, P., Slooten, E., & Dawson, S. M. (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*, **23**(4), 360–371.

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

Meester, R., & Roy, R. (1996). *Continuum percolation*. Cambridge, UK: Cambridge University Press.

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

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. (2002). Assortative mixing in networks. *Physical Review Letters*, **89**(20), 208701.

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

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

Newman, M. E. J. (2003c). Random graphs as models of networks. In S. Bornholdt, & H. G. Schuster (Eds.), *Handbook of graphs and networks* (35–68). Berlin: Wiley-VCH.

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. (2009). Random graphs with clustering. *Physical Review Letters*, **103**(5), 058701.

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

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

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

Olhede, S. C., & Wolfe, P. J. Degree-based network models. *arxiv:1211.6537*.

Padgett, J. F., & Ansell, C. K. (1993). Robust Action and the Rise of the Medici, 1400–1434. *American journal of sociology*, **98**(6), 1259–1319.

Penrose, M. D. (1991). On a continuum percolation model. *Advances in Applied Probability*, **23**, 536–556.

Perry, P. O., & Wolfe, P. J. (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., Niu, X., Hoff, P. D., & Yeung, K. Y. (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*. Ph.D. thesis, Cornell University, September.

Schweinberger, M., & Handcock, M. S. (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., & Rinaldo, A. (2013). Consistency under sampling of exponential random graph models. *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., Thomas, A. C., & Junker, B. W. (2013). Hierarchical network models for education research: Hierarchical latent space models. *Journal of Educational and Behavioral Statistics*, **38**(3), 295–318.

Wang, H., Tang, M., Park, Y., & Priebe, C. E. (2014). Locality statistics for anomaly detection in time series of graphs. *IEEE Transactions on Signal Processing*, **62**, 703–717.

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

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

https://doi.org/10.1017/nws.2016.23 Published online by Cambridge University Press