On the geometry of discrete exponential families with application to exponential random graph models

Alessandro Rinaldo

Department of Statistics
Carnegie Mellon University
e-mail: arinaldo@stat.cmu.edu

Stephen E. Fienberg

Department of Statistics,
Machine Learning Department and Cylab
Carnegie Mellon University
e-mail: fienberg@stat.cmu.edu

and

Yi Zhou[*]

Machine Learning Department
Carnegie Mellon University
e-mail: yi@stat.cmu.edu

Abstract: There has been an explosion of interest in statistical models for analyzing network data, and considerable interest in the class of exponential random graph (ERG) models, especially in connection with difficulties in computing maximum likelihood estimates. The issues associated with these difficulties relate to the broader structure of discrete exponential families. This paper re-examines the issues in two parts. First we consider the closure of k-dimensional exponential families of distribution with discrete base measure and polyhedral convex support P. We show that the normal fan of P is a geometric object that plays a fundamental role in deriving the statistical and geometric properties of the corresponding extended exponential families. We discuss its relevance to maximum likelihood estimation, both from a theoretical and computational standpoint. Second, we apply our results to the analysis of ERG models. By means of a detailed example, we provide some characterization of the properties of ERG models, and, in particular, of certain behaviors of ERG models known as degeneracy.

---

[*]Current affiliation: Barclays Bank, New York.

1. Introduction

Our motivation for the work described in this paper comes from the analysis of network data using models representable by graphs, where the nodes correspond to individuals and the edges to relations or linkages among them. Such graphical representation has a long history, dating back to Moreno (1934), and was recast within the exponential family framework by Holland and Leinhardt (1981) and Frank and Strauss (1986) (see also Strauss and Ikeda, 1990). Their work led to the development of the broader class of exponential random graph (ERG), or p^*, models for social networks (see, e.g. Wasserman and Pattison, 1996), but likelihood methods for their analysis remained out of reach until earlier this decade. For a broad review of these and other network models, see Goldenberg et.al. (2009). Recent work on maximum likelihood estimation for ERG models, however, has pointed to difficulties that have been characterized as "degeneracies" or "near degeneracies" by Handcock (2003) and Hunter et al. (2008). The explanation for these difficulties lies within broader characterizations of "degeneracies" for discrete exponential families.

Exponential families are one of the most important and widespread class of parametric statistical models, whose remarkable properties have long been established in the statistical literature (see, e.g., Bardoff-Nielsen, 1978; Brown, 1986; Letac, 1992). Among the most interesting features of exponential families is the notion of the closure of the family, known as the extended exponential family, whose mathematical theory has been recently worked out in great generality (see Csiszár and Matúš, 2001, 2003, 2005, 2008). The study of the extended families is particularly important, as it may directly pertain to the existence of the maximum likelihood estimates and to the estimability of the natural parameters. This is the case for discrete exponential families, for which the maximum likelihood estimates may not exist with some positive probability. A notable instance is the class of log-linear models, for which existence of the MLE and closure of the family can be characterized in a purely geometric fashion (see, e.g., Eriksson et al., 2006; Geiger et al., 2006; Rinaldo, 2006a).

In this article we are concerned with discrete linear exponential families. In the first part of the paper, we show that the geometric and statistical properties of the extended family depend in a fundamental way on the normal fan of the convex support. In particular, the normal fan can be used to characterize non-identifiability of the families in the closure, to represent the densities in the extended family as almost sure limits of the densities in the original family along certain directions of the parameter space and to describe the directions of recession of the (negative) log-likelihood function.

As an application of our results, in the second part of the paper we turn our attention to exponential random graph models, a particular class of discrete linear exponential families. Our discussion is based an the detailed analysis of the simple ERG model on the graphs on 9 nodes with two-dimensional natural sufficient statistics consisting of the number of edges and the number of triangles that we learned from Handcock (2003). We use Shannon's entropy function to illustrate graphically how concentrated the distributions in this family are, viewed

as functions of both the natural and mean value parameters. Besides illustrating the theoretical results derived in the first part of the article, our analysis sheds some light on a variety of pathological behaviors observed in practice while fitting some ERG models known as degeneracy (see, e.g., Handcock, 2003), and, we believe, more generally, on the qualities and attributes of ERG models. Our study suggests that degeneracy should not be regarded as a set of pathological features, rather it can be explained by the properties of exponential families with polyhedral support. Although our conclusions and observations reflect a somewhat common perception in the social network community, our arguments are rigorous.

The remainder of this article is organized as follows. In section 2 we provide the derivation of our key theoretical results. In section 2.1, we begin by describing our settings and briefly review the theory of extended exponential families and their fundamental properties. In section 2.2, we introduce the notions of normal cones and the normal fan to the convex support of the family. In section 2.3 we state our main result and a discussion of its corollaries, while section 2.4 presents come computational considerations concerning maximum likelihood estimation for extended exponential families. Section 3 consists of an application of our results to ERG models. First in section 3.1 we introduce the class of ERG models and then in section 3.2 we present our running example of an ERG model on the set of all graphs on 9 nodes. We next introduce the concept of degeneracy for ERG models in section 3.3, while in section 3.4 we use our theoretical results to illustrate graphically the features of the model in the running example of section 3.2 to show how degeneracy arises. The appendices contain the proofs and some additional results on how to establish existence of the maximum likelihood estimates in discrete linear exponential families using linear programming.

2. Extended exponential families with polyhedral support

We use the following notation throughout this section and the remainder of the paper:

- For two vectors x and y in $\mathbb{R}^d$, denote their *inner product* by $\langle x, y \rangle = \sum_{i=1}^d x_i y_i$.
- The *Euclidean norm* of a vector x is $\|x\|_2 = \sqrt{\langle x, y \rangle}$.
- For any $A \subset \mathbb{R}^d$, we denote with convhull(A) its *convex hull*, i.e. the set of all convex combinations of points in A; similarly we denote with cone(A) its *conic hull*, i.e. the set of all of conic combinations of points in A.
- For any $A \subset \mathbb{R}^d$, we denote its *relative interior* and *relative boundary* with respect to the affine space spanned by A, by ri(A) and rb(A), respectively.

2.1. Statistical and geometric background

In this section we introduce the statistical and geometric background needed for our results and review the basic concepts from the general theory of exponential

families and polyhedral geometry. For more complete treatments we refer the reader to Bardoff-Nielsen (1978), Brown (1986) and Csiszár and Matúš (2001, 2003, 2005, 2008) for material on exponential families, and Ziegler (1996) and Schrijver (1998) for introductions to polyhedral geometry.

We consider an exponential family of distributions $\mathcal{E}_P$ on $\mathbb{R}^k$ with densities

$$p_\theta(x) = \exp\left\{\langle x, \theta \rangle - \psi(\theta)\right\}, \quad \theta \in \Theta,$$

with respect to some σ-finite base measure ν, where

$$\Theta \subseteq \{\theta \in \mathbb{R}^k \colon \int_{\mathbb{R}^k} \exp^{\langle x, \theta \rangle} d\nu(x) < \infty\}$$

is the natural parameter space and $\psi(\theta) = \log \int_{\mathbb{R}^k} \exp^{\langle x, \theta \rangle} d\nu(x)$ the log-partition function. The support of $\mathcal{E}_P$ is the closure of the set $\{x \colon \nu(x) > 0\}$, while the *convex support* P is the closure of the convex hull of the support of $\mathcal{E}_P$. We will assume throughout that

(A1) ν has countable support;
(A2) P is a full-dimensional polyhedron in $\mathbb{R}^k$, that is, P does not belong to any proper affine subspace of $\mathbb{R}^k$;
(A3) for each face F of P, $F = \mathrm{convhull}(S_F)$, for some set $S_F \subset \mathrm{supp}(\nu)$;
(A4) the natural parameter space Θ is an open set in $\mathbb{R}^k$.

Assumptions (A1) and (A2) imply, in particular, that the family is in minimal form and, therefore, identifiable. We remark that full-dimensionality of P in assumption (A2) is not necessary and is imposed to simplify the exposition; our results would still hold with some minor changes, at the cost of additional technicalities in the proofs. In fact, any degenerate exponential family can be made full by taking appropriate affine transformations, a procedure known as reduction to minimality (see, e.g., theorem 1.9 in Brown, 1986 or lemma 8.1 in Bardoff-Nielsen, 1978). Assumptions (A1) and (A3) are needed to guarantee the existence of probability distributions supported over the boundary of P, which is an indispensable feature of the extended exponential family, described in the next section. Condition (A3) could be relaxed by allowing some faces to have zero ν measure. Finally, assumption (A4) is a standard. In particular, for our discussion of ERG models, $\Theta = \mathbb{R}^k$.

Assumption (A2) means that P can be represented as the intersection of a finite number of half-spaces. Formally, there exists a $m \times k$ matrix A and a vector $b \in \mathbb{R}^m$ such that

$$P = \{x \in \mathbb{R}^k \colon Ax \leq b\}, \tag{1}$$

where the system contains no implicit equalities (meaning that no inequality is in fact an equality satisfied by all the points in P). A proper face F of P is a subset of P defined by

$$F = \{x \in P \colon A_F x = b_F\}, \tag{2}$$

for some subsystem $A_F x \leq b_F$ of $Ax \leq b$ and, therefore, it is itself a polyhedron. The whole polyhedron P is regarded as the improper face of itself associated to

the full system of inequalities, so that P is representable as the disjoint union of the relative interiors of all its faces. The dimension of a face F, $\dim(F)$, is the dimension of the affine subspace it generates or, equivalently, the dimension of the null space of A_F. Faces of dimension $k-1$ are called facets of P and, if the system (1) has no redundant inequality, something which can always be assumed without loss of generality, the number m of rows of A matchs the number of facets. Equation (1) is known as the $\mathcal{H}$-representation of P. Alternatively, P could be described using the $\mathcal{V}$-representation as

$$P = Q + C, \tag{3}$$

where the symbol $+$ denotes the Minkowski sum, $Q = \text{convhull}(\mathcal{Q})$ and $C = \text{cone}(\mathcal{C})$, with $\mathcal{Q}$ and $\mathcal{C}$ two finite sets of vectors in $\mathbb{R}^k$. The Minkowski sum of two subsets A and B of $\mathbb{R}^k$ is the set $A + B = \{x + y, x \in A, y \in B\} \subset \mathbb{R}^k$. If, in the $\mathcal{V}$-representation of P, $C = \{0\}$, then P is a bounded polyhedron, or a polytope; similarly, if $Q = \{0\}$, P is a polyhedral cone, an unbounded set. Throughout the paper, we will typically rely on the $\mathcal{H}$-representation (1), which we find more suited to our purposes, although our results could be established using (3).

2.1.1. Basics of extended exponential families

Let $x \in \mathbb{R}^k$ be the observed sample corresponding to a random vector X having an unknown distribution in $\mathcal{E}_{\mathrm{P}}$, which we will refer to as the natural sufficient statistics. The random set

$$\widehat{\theta}(x) = \widehat{\theta} = \left\{ \theta^* \in \Theta \colon p_{\theta^*}(x) = \sup_{\theta \in \Theta} p_\theta(x) \right\} \tag{4}$$

is the *maximum likelihood estimate*, or MLE, of θ. If $\widehat{\theta} = \emptyset$ the MLE is said to be nonexistent. Existence of the MLE is determined by the geometry of P, as indicated by the following well-known, fundamental result (see, e.g., theorem 5.5 in Brown, 1986 or proposition 4.2 Rinaldo, 2006a for different proofs).

Theorem 2.1. *Under conditions (A1)–(A4), the MLE $\widehat{\theta}$ exists and is unique if and only if $x \in \text{ri}(\mathrm{P})$.*

Setting $\mathbb{E}_\theta(X) = \int_{\mathbb{R}^k} z p_\theta(z) d\nu(z)$, because of the minimality of $\mathcal{E}_\mathcal{P}$, one can obtain the fundamental identity

$$\nabla\psi(\theta) = \mathbb{E}_\theta(X), \tag{5}$$

valid for each $\theta \in \Theta$, where ∇ indicates the gradient. The map $\nabla\psi\colon \Theta \mapsto \text{ri}(\mathrm{P})$ given by (5) is, in fact, a homeomorphism, so that one can equivalently represent any distribution in $\mathcal{E}_\mathrm{P}$ using the natural parameter θ or the mean value parameter $\mu = \mu(\theta) = \mathbb{E}_\theta(X) \in \text{ri}(\mathrm{P})$. This alternative parametrization is

is known as the *mean value parametrization*. In particular, if the MLE exists, it is determined by the equation

$$\widehat{\theta} = (\nabla\psi)^{-1}(x),$$

which translates into the moment equation $\mathbb{E}_{\widehat{\theta}}(X) = x$.

For any proper face F, let ν_F be the restriction of ν to F. Then, ν_F determines a new exponential family of distributions $\mathcal{E}_F$, with densities with respect to ν_F given by

$$p_\theta^F(x) = \exp\left\{\langle x, \theta\rangle - \psi^F(\theta)\right\}, \quad \theta \in \Theta_F,$$

where the natural parameter space is $\Theta_F = \{\theta \in \Theta \colon \int_{\mathbb{R}^k} \exp^{\langle x,\theta\rangle} d\nu_F(x) < \infty\}$ and the log-partition function is $\psi^F(\theta) = \log\int_{\mathbb{R}^k} \exp^{\langle x,\theta\rangle} d\nu_F(x)$. Notice that, since $\int_{\mathbb{R}^k} \exp^{\langle x,\theta\rangle} d\nu_F(x) \leq \int_{\mathbb{R}^k} \exp^{\langle x,\theta\rangle} d\nu(x)$, $\Theta = \Theta_F$. By assumption (A3), the convex support of this new family is F and the existence result of theorem 2.1 carries over: the MLE exists if and only if the observed sample x belongs to $\mathrm{ri}(F)$. Since $\mathcal{E}_F$ is supported on a lower-dimensional affine subspace of $\mathbb{R}^k$, however, it is no longer minimal. Hence, the MLE is not unique, and it consists instead of many solutions to (4); see corollary 2.9 below for details. Nonetheless, via reduction to minimality (see, e.g., Brown, 1986, theorem 1.9), we can verify that, when $\widehat{\theta}$ is not empty, it consists exactly of those θ satisfying the first order optimality conditions

$$x = \nabla\psi_F(\theta), \tag{6}$$

with the corresponding moment equations $\mathbb{E}_\theta^F(X) = \int_{\mathbb{R}^k} z p_\theta^F(z) d\nu^F(z) = x$ still holding, $\forall \theta \in \widehat{\theta}$. In fact, lack of minimality bears no effect on the mean value parametrization: for every $\theta \in \Theta_F$, there exists one point $x \in \mathrm{ri}(F)$ such that

$$\mathbb{E}_\theta(X) = x, \tag{7}$$

and, similarly, for any $x \in \mathrm{ri}(F)$, there exists a set $\theta_F \subset \Theta_F$, depending on x, such that (7) holds for all $\theta \in \theta_F$. See equation (10) below for a characterization of θ_F.

The collection of distributions

$$\mathcal{E} = \bigcup_F \mathcal{E}_F$$

as F ranges over all the faces of P, including P itself, is known as the *extended exponential family* of distributions. With respect to the extended family $\mathcal{E}$, for *any* observed sample x, the MLE, or *extended MLE*, is always well defined and is the set of solutions to (6), where F is the *unique* face containing x in its relative interior.

2.2. *Extended exponential families and the normal fan of* P

In this section we introduce the notion of normal fan of P and establish its relevance for the extended family $\mathcal{E}$. See lemma B2 in Appendix B for some basic properties of the normal cones and of the normal fan.

For every face F of P, let

$$N_F = \left\{ c \in \mathbb{R}^k \colon F \subseteq \{x \in \mathrm{P} \colon \langle c, x \rangle = \max_{y \in \mathrm{P}} \langle c, y \rangle \} \right\}$$

be the polyhedral cone consisting of all the linear functionals on P that are maximal over F, known as the *normal cone* of F. Then, $\dim(N_F) = k - \dim(F)$, so that larger faces of P correspond to smaller normal cones. By lemma B2 part 5., the normal cone of a proper face F can be equivalently defined as

$$N_F = \mathrm{cone}\,(a_1, \ldots, a_{m_F}),$$

where a_i denotes the transpose of the i-th row of the submatrix A_F given in (2), $i = 1 \ldots, m_F$.

The collection of cones

$$\mathcal{N}(\mathrm{P}) = \{N_F, F \text{ is a face of P}\}$$

forms a polyhedral complex in $\mathbb{R}^k$ (see, e.g. Sturmfels, 1995), known as the *normal fan* of P. Notice that, since $\dim(\mathrm{P}) = k$, $N_\mathrm{P} = \{0\}$ and $\mathcal{N}(P)$ is pointed, which implies that none of the normal cones contains a linear subspace. Furthermore,

$$\biguplus_{N_F \in \mathcal{N}(P)} \mathrm{int}(N_F) = \mathrm{C}^*,$$

where $C^* = \{x \in \mathbb{R}^k \colon \langle x, y \rangle \leq 0, \forall y \in \mathrm{C}\}$ is the polar of C in the $\mathcal{V}$-representation (3) of P and $\biguplus$ denotes disjoint union. In particular, if $\mathrm{C} = \{0\}$, i.e. if P is a full-dimensional polytope, the cones in $\mathcal{N}(\mathrm{P})$ partition $\mathbb{R}^k$:

$$\biguplus_{N_F \in \mathcal{N}(P)} \mathrm{int}(N_F) = \mathbb{R}^k. \tag{8}$$

We mention that, more generally, if full-dimensionality in assumption (A2) is not in force, then N_P is a linear subspace of $\mathbb{R}^k$ of codimension $k - \dim(\mathrm{P})$.

Let $\mathrm{lin}(N_F)$ denote the subspace generated by N_F, which is the linear subspace spanned by the vectors $(a_1, \ldots, a_{m_F})$. The following lemma shows that, for every face F of the convex support, the parameter space of the extended family $\mathcal{E}_F$ can be fully described using $\mathrm{lin}(N_F)$. Below, for two linear subspaces S_1 and S_2 of $\mathbb{R}^k$, the quotient space of S_1 modulo S_2 is the vector subspace $\{[x], \colon x \in S_1\}$ consisting of the equivalence classes on S_1 of the form $[x] = \{y \in S_1 \colon x - y \in S_2\}$.

Lemma 2.2. *The family $\mathcal{E}_F$ is non-identifiable and Θ_F is the quotient space of Θ modulo $\mathrm{lin}(N_F)$. Furthermore, for any $\zeta \in \mathrm{lin}(N_F)$,*

$$\mathrm{rank}\,(I_F(\theta + \zeta)) = \dim(F), \tag{9}$$

where $I(\cdot)$ and $I_F(\cdot)$ denote the Fisher information matrices for $\mathcal{E}_\mathrm{P}$ and $\mathcal{E}_F$, respectively.

The previous result characterizes Θ_F as the vector subspace of equivalence classes of points in Θ, where θ_1 and θ_2 are in the same class if and only if $\theta_1 - \theta_2 \in \mathrm{lin}(N_F)$, and the class containing $\theta \in \Theta$ is the set

$$\theta_F \equiv \{\theta + \zeta \in \Theta, \zeta \in \mathrm{lin}(N_F)\}. \tag{10}$$

Notice that if $\Theta = \mathbb{R}^k$, then Θ_F is comprised of affine subspaces of dimension $\dim(N_F) = k - \dim(F)$ parallel to $\mathrm{lin}(N_F)$, each identifying a single distribution. In particular, when $F = \mathrm{P}$, $\mathrm{lin}(N_F) = \{0\}$, so that θ_F is an atomic set and we recover the original, fully identifiable family $\mathcal{E}_\mathrm{P}$.

By lemma 2.2, for any $\theta \in \theta_F$, p_θ^F is the same density. For the remainder of the article, we will denote this density with $p_{\theta_F}^F$; similarly, we will write $\mathbb{E}_{\theta_F}(X) = \int_{\mathbb{R}^k} z p_{\theta_F}^F(z) d\nu_F(z)$.

2.3. Main result

We will utilize the normal fan $\mathcal{N}(\mathrm{P})$ to describe various cases in which the following convergence statements will hold or fail:

$$p_{\theta_n} \to p_{\theta_F}^F, \ a.e. \ \nu, \quad \text{and} \quad \mu_n \to \mu_F \in \mathrm{ri}(F), \tag{11}$$

where $\{\theta_n\}$ is a sequence of parameters in Θ, $\mu_n = \mathbb{E}_{\theta_n}(X)$, $\mu_F = \mathbb{E}_{\theta_F}(X)$, and F is any *proper* face of P. Characterizing the cases in which (11) is verified is of relevance as these statements explicitly provide equivalent representations of the extended family $\mathcal{E}$ as the closure of the original family $\mathcal{E}_\mathrm{P}$ in both natural and mean value parameterization, and also in terms of almost sure limits of the densities in $\mathcal{E}_\mathrm{P}$. Thus, the geometric analysis of (11) we are about to present is a more powerful representation of $\mathcal{E}$ than the traditional construction described in section 2.1.1. We take note that, because of the one-to-one correspondence between natural and mean value parameters for the families comprising $\mathcal{E}$, the two conditions in (11) imply each other. Of course, if $F = \mathrm{P}$, this is just a restatement of the mean value parametrization homeomorphism (5) between Θ and $\mathrm{ri}(\mathrm{P})$.

As a preliminary observation, we point out that (11) holds true only if the parameters θ_n have diverging norms, so that $p_{\theta_F}^F$ cannot belong to $\mathcal{E}_\mathrm{P}$. Formally,

Lemma 2.3. *If (11) is verified for some proper face F, then $\|\theta_n\|_2 \to \infty$.*

In our main result, we establish sufficient conditions under which (11) holds or fails, based on the cones in the normal fan of P.

Theorem 2.4. *Assume that conditions (A1)–(A4) hold. Let $\{\theta_n\} \subset \Theta$ be a sequence of natural parameters satisfying $\theta_n = \eta + \rho_n d_n$, where $\{\rho_n\}$ is a sequence of non-negative scalars tending to infinity, $\eta \in \theta_F \cap \Theta$ and $\{d_n\}$ is a sequence of unit vectors.*

1. *If $\{d_n\} \subset R$, with R a compact subset of $\mathrm{ri}(N_F)$, then (11) holds*
2. *Conversely, if $\{d_n\} \subset R$, with R a compact subset of N_F^c, then (11) fails.*

3. If $\{d_n\} \subset R$, with R a compact subset $(\mathcal{N}(\mathrm{P}))^c$, then

$$\|\mu_n\|_2 \to \infty, \tag{12}$$

which, in particular, implies that (11) is not verified.

Remark

1. The condition $\|d_n\|_2 = 1$ for all n is imposed to guarantee that $\|\rho_n d_n\|_2 \to +\infty$ and it does not entail any loss in generality.
2. The theorem shows that (11) will hold or fail *uniformly* over compact subsets of $\mathrm{ri}(N_F)$, for all faces F of P.

Below, we will concern ourselves with sequences $\{\theta_n\}$ of natural parameters of a certain simplified form, as described below.

Definition 2.5. A sequence of natural parameters $\{\theta_n\} \subset \Theta$ is a $(\theta, d, \{\rho_n\})$-sequence if

$$\theta_n = \theta + \rho_n d,$$

where $\theta \in \Theta$, $d \in \mathbb{R}^k \setminus \{0\}$ and $\{\rho_n\}$ is a sequence of non-negative numbers tending to infinity.

The restriction to $(\theta, d, \{\rho_n\})$-sequences is a strong enough condition to yield a full characterization of (11), as described in the next corollary, and yet sufficiently mild to unveil some of the fundamental features of the extended family $\mathcal{E}$. Furthermore, it will allow us to recast some of our results in the language of convexity theory and gain some insights on the computational aspects of calculating the extended MLE.

Corollary 2.6. *Let $\{\theta_n\}$ be a $(\theta, d, \{\rho_n\})$-sequence.*

1. *The convergence statements in (11) hold if and only if $d \in \mathrm{ri}(N_F)$.*
2. *If $d \notin \mathcal{N}(\mathrm{P})$, then (12) is verified.*

In essence, corollary 2.6 characterizes the extended family $\mathcal{E}$ as the compactification of the original family $\mathcal{E}_\mathrm{P}$ under both natural and mean value parametrizations. For the natural parametrization, each density in $\mathcal{E}_F$ is obtained as the point-wise limit of sequences of densities parametrized by sequences of points in Θ along any direction in $\mathrm{ri}(N_F)$ with norms diverging to infinity. In contrast, the corresponding sequence of mean value parameters converges to the corresponding point of finite norm on the boundary of P. This is a striking difference between natural and mean value parametrization, which is entirely captured by the normal fan of P. See figures 3 and 4 in section 3.4 and related discussion for more details in the context of ERG models. See also the short movies available at http://www.stat.cmu.edu/~arinaldo/ERG/ for a more intuitive graphical illustration.

In the remaining of this section, we will explore some of the consequences of theorem 2.4 and, in particular, of corollary 2.6, with the goal of illustrating some of the key properties of the extended family $\mathcal{E}$.

We begin by observing that, as shown in equation (20) in the proof of theorem 2.4, if $d \notin N_F$, the sequence of distributions parametrized by the points $\theta_n = \theta + \rho_n d$ corresponds to distributions in the original family $\mathcal{E}_P$ whose mean value parameters $\mu_n \in \mathrm{ri}(P)$ are such that $\|\mu_n\|_2 \to \infty$, with μ_n bounded away from $\mathrm{rb}(P)$. It is clear that this can occur only if $C \neq \{0\}$, i.e. if the convex support is unbounded. In fact, when P is a polytope, we have $\mathcal{N}(P) = \mathbb{R}^k$ (see equation 8), so that corollary 2.6 further yields that each density in the family $\mathcal{E}_F$ can be obtained as $\lim_n p_{\theta_n}$, where $\{\theta_n\}$ is any $(\theta, \{\rho_n\}, d)$-sequence with $\theta \in \Theta$ and $d \in \mathrm{ri}(N_F)$. Formally,

Corollary 2.7. *If P is a polytope, then, for any* $d \in \mathbb{R}^k$*, any* $(\theta, \{\rho_n\}, d)$*- sequence* $\{\theta_n\}$ *and any face F,*

$$p_{\theta_n} \to p^F_{\theta_F}, \ a.e. \ \nu, \quad and \quad \mu_n \to \mu_F \in \mathrm{ri}(F),$$

if and only if $d \in \mathrm{ri}(N_F)$*.*

In fact, our analysis of exponential random graph models of section 3.4 is almost entirely an illustration of the previous corollary.

Another implication of corollary 2.6 is that, when the MLE does not exist, the directions of increase of the likelihood function for a given observed sample $x \in \mathrm{ri}(F)$ are precisely the points in the associated normal cone N_F. Formally, let x be the observed natural sufficient statistics and $\ell_x : \Theta \mapsto \mathbb{R}$ be the log-likelihood function, given by $\ell_x(\theta) = \log p_\theta(x)$. Then, $-\ell_x$ is a strictly convex function. This follows from minimality and the well-known convexity properties of the cumulant generating function ψ (see, e.g., Brown, 1986, theorem 1.13). Then, following Rockafellar (1970, Chapter 8), $d \in \mathbb{R}^k \setminus \{0\}$ is a direction of recession for $-\ell_x$ if

$$\liminf_{\rho \to \infty} - \ell_x(\theta + \rho d) < \infty, \tag{13}$$

for one, and thus for all, $\theta \in \mathrm{dom}(\ell_x) = \Theta$. The set of all directions of recession of $-\ell_x$, along with the point $d = 0$, is called the recession cone of $-\ell_x$. It is clear that convex functions admitting directions of recession might not achieve their infimum at any point in their effective domain. On account of the next result, the recession cone of $-\ell_x$ is a cone of the normal fan of P, almost everywhere ν.

Corollary 2.8. *For any observable natural sufficient statistics* $x \in P$*, the poly-hedral cone* N_F *is the recession cone of the negative log-likelihood function* $-\ell_x$*, where F is the unique, possibly improper, face of P such that* $x \in \mathrm{ri}(F)$*.*

In particular, when $x \in \mathrm{ri}(P)$, i.e. when the MLE exists, the corresponding recession cone is just the point $\{0\}$ (since $\dim(P) = k$), so that the negative log-likelihood function does not have any direction of recession and, therefore, its infimum is achieved at one parameter point $\widehat{\theta} \in \mathbb{R}^k$ with finite norm, namely the MLE. On the other hand, when the MLE is nonexistent, the likelihood function increases for any sequence of natural parameters with norm diverging to infinity along any nonzero direction $d \in N_F$, where N_F is the normal cone of the face of P containing the observed natural sufficient statistics in its relative interior.

During the preparation of the paper, we learned of similar results in Geyer (2008), which are based on the characterization of the convex support in term of the tangent cones and normal cones. While his analysis applies to more general classes of exponential families, our results are more refined, as we take full advantage of the polyhedral assumption and establish a direct connection between the extended families and the cones in the normal fan.

By combining the results derived so far, we next show that, when $x \in \mathrm{ri}(F)$, the extended MLE will be the affine subspace of dimension $\dim(N_F)$ given by $\widehat{\theta}_F$, where $\mathbb{E}_{\widehat{\theta}_F} = x$. Though not entirely a new result (see Brown, 1986, Chapter 6), our proof and the characterization of $\widehat{\theta}_F$ in terms of N_F is novel.

Corollary 2.9. *Let* $x \in \mathrm{ri}(F)$ *and* $\widehat{\theta}_F$ *be given as in (10) and such that* $\mathbb{E}_{\widehat{\theta}_F}(X) = x$. *Then,*

$$\sup_{\theta \in \Theta} p_\theta(x) = p^F_{\widehat{\theta}_F}(x).$$

For completeness, we conclude this section by linking our discussion with alternative characterizations of the closure of the family $\mathcal{E}_\mathrm{P}$ existing in the literature, which could be easily obtained using theorem 2.4 (see, in particular, Csiszár and Matúš, 2001, 2003, 2005, 2008).

Corollary 2.10. *For any* $(\theta, \{\rho_n\}, d)$*-sequence* $\{\theta_n\}$ *with* $d \in \mathrm{ri}(N_F)$,

   i) $P_{\theta_n} \overset{\mathrm{TV}}{\to} P^F_{\theta_F}$, *where* $\overset{\mathrm{TV}}{\to}$ *denotes convergence in total variation;*

  ii) $\lim_n K(P^F_{\theta_F}, P_{\theta_n}) = 0$, *where* $K(P, Q)$ *is the Kullback-Lieber divergence of* P *from* Q;

 iii) $P_{\theta_n} \Rightarrow P^F_{\theta_F}$, *where the* $\Rightarrow$ *denotes convergence in distribution.*

2.4. Computational considerations

Based on our findings, we can make a few observations regarding the computational difficulties of finding the extended MLE, some of which are exemplified in the next result.

Corollary 2.11. *Let* $\{\theta_n\}$ *be a* $(\theta, \{\rho_n\}, d)$*-sequence, with* $d \in \mathrm{ri}(N_F)$. *Then, for every* $\zeta \in \mathrm{lin}(N_F)$,

$$I(\theta_n) \to I_F(\theta + \zeta), \tag{14}$$

where convergence is pointwise.

From this corollary and equation (9), we can infer that, when the MLE does not exist, maximizing the log-likelihood function using the Newton-Raphson method, as well as virtually any other ascent methods, may fail due to numerical instabilities. In fact, the Newton-Raphson algorithm proceeds by finding a sequence $\{\theta_n\}$ of natural parameters along which ℓ_x increases; at each step of the procedure, the next point in the sequence is determined along the direction specified by the inverse of the Hessian, i.e. by the inverse of $I(\theta_n)$, times the gradient of $p_\theta(x)$. (These are directions of fastest ascent measured in the Fisher

information norm). However, for all n large enough, the Hessian matrices will be badly conditioned, since, at the optimum, the Fisher information matrix is not invertible (see equation 9). These singularity issues become more severe if the observed statistics x belong to the relative interior of a face of small dimension. Furthermore, unless x lies on a the relative interior of a facet, there is an infinite number of directions along which the likelihood function increases. It is apparent that these problems are even more common in high-dimensional settings or whenever the data are sparse.

From the statistical standpoint, equations (14) and (9) further imply that, when the MLE does not exist, the standard error may be quite large, and that the number of degrees of freedom should be adjusted to reflect the non-estimability of some parameters. As a result, any hypothesis testing or model selection procedure that rely solely on these estimates should be regarded, at the very least, unreliable. Based on these considerations, it is clear that not only is the task of computing the extended MLE particularly daunting, but the statistical interpretation of these quantities is also rather delicate.

We refer the reader to Geyer (2008) and Rinaldo (2006b) for different algorithmic approaches to computing the extended MLE for certain types of exponential models with polyhedral support for which a $\mathcal{V}$-representation of P of the form (1) or (3) is either available or easily computable. We remark that, in order to determine the extended MLE, it is necessary not only to have an explicit representation of P but, in addition, to be able to have in closed form the log-partition functions ψ^F, for each face F. A class of models for which both conditions are satisfied is the class of the log-linear models. If this information is not available, one may resort to MCMC techniques for computing the MLE or a pseudo-MLE, as for the class of models to be described in the next section. See Geyer and Thompson (1992), and Handcock (2003), Snijders (2002), Wasserman and Robins (2004), Handcock et al. (2006), Hunter et al. (2008) and references therein.

As a final comment, we point out that, while computing the extended MLE is very often a hard problem, deciding whether the MLE exists is typically more feasible, and can be accomplished using linear programming, provided an explicit representation, namely a $\mathcal{H}$- or a $\mathcal{V}$-representation, of P is available. See Appendix C for details and also Eriksson et al. (2006) for an application to hierarchical log-linear models.

3. Application to exponential random graph models

We now apply some of the results from the previous section to the class of exponential random graph models. The motivation for our choice is the attempt to explain certain features of ERG models that have been observed empirically and have been collectively labeled as degeneracy (see, e.g., Handcock, 2003). Our point of view is simply that there is nothing degenerate or unusual about these models, whose behavior can in fact be explained in a direct way using the properties of exponential families with polyhedral support as described in section 2.3.

TABLE 1
Some information about the complexity of some ERG models on small graphs

Number of nodes: g	Possible number of edges: $\binom{g}{2}$	Number of graphs: $|\mathcal{G}_g|$
7	21	$2,097,152$
8	28	$268,435,456$
9	36	$68,719,476,736$
10	45	$35,184,372,088,832$

Our arguments rely on a thorough analysis of one ERG model, described below in section 3.2, and on graphical renderings of corollary 2.7, which we find particularly effective and elucidative of our results. We looked at a variety of other ERG models on 7,8 and 9 nodes, using different choices of the network statistics described below, and arrived to the same kind of conclusions we are about to present.

Finally, we would like to emphasize that, as the log-partition function is not available in closed form, an exact analysis of ERG models on larger graphs is almost impossible. This is due to the need to enumerate all possible graphs with a given number of nodes in order to evaluate that function, a task whose computational computationally becomes prohibitive very rapidly as the number nodes grow; see equation (15) below and Table 1.

3.1. Introduction to ERG models

There is an extensive literature on ERG models and their use in social network analysis. A partial but representative list of references is: Holland and Leinhardt (1981), Frank and Strauss (1986), Wasserman and Pattison (1996), Wasserman and Robins (2004), Robins et al. (2007a,b) and references therein. Below we briefly describe the settings for ERG models, in order to make explicit the connections with the material in the previous sections.

Consider the set $\mathcal{G}_g$ of all possible simple, i.e. unweighted, undirected and without loops, graphs on g nodes. Every such graph x can be described by a 0-1 symmetric $g \times g$ adjacency matrix, whose (i,j)-th entry is 1 if there exists an edge between the nodes i and j and 0 otherwise. Thus, x can be represented as a $\binom{g}{2}$-dimensional 0-1 vector. The cardinality of $\mathcal{G}_g$ grows super-exponentially in the number of nodes, namely

$$|\mathcal{G}_g| = 2^{\binom{g}{2}}, \tag{15}$$

so that network modeling entails constructing probability distributions over very large discrete spaces (see Table 1).

Let $T \colon \mathcal{G}_g \mapsto \mathbb{R}^k$ be a vector valued function of *network statistics* quantifying the key features of interest of a given observed graph. In this article we are mostly concerned with ERG models arising from network statistics that capture rather general and aggregate features of the network. Typical examples of such statistics are (see, e.g., Goodreau, 2007, for more details):

1. the number of edges: $\sum_{i<j} x_{ij}$
2. the number of triangles: $\sum_{i<j<h} x_{ij} x_{jh} x_{ih}$
3. the k-degree statistic: $D_k(x) = \sum_{i=1}^{g} 1\{d_i = k\}$, where $d_i = \sum_j x_{ij}$ is the degree of the i-th node and $0 \le k \le n-1$;
4. the number of k-stars: $\sum_{i=k}^{g-1} \binom{i}{k} D_i(x)$, $2 \le k \le n-1$, i.e. the number of distinct edges that are incident to the same node, where $D_i(x)$ is the i-th degree statistic given above;
5. the alternating k-star statistic

$$\sum_{i=2}^{g-1} (-1)^{i-1} \frac{S_i(x)}{\lambda^{2-i}},$$

where λ is a positive parameter.

For all modeling purposes, these network statistics are effectively regarded as sufficient statistics and, by the Koopman-Pitman-Darmois theorem (see, e.g. Brown, 1986; Bardoff-Nielsen, 2006), the resulting exponential family of distributions provides a convenient statistical model for $\mathcal{G}_g$. Formally, given a set of network statistics in the form of a k-valued function $T(\cdot)$ on $\mathcal{G}_g$, the ERG model $\mathcal{P} \equiv \{Q_\theta, \theta \in \Theta \subseteq \mathbb{R}^k\}$ is the exponential family of probability distributions over $\mathcal{G}_g$ with natural sufficient statistics $T(x)$ and base measure μ given by the counting measure on $\mathcal{G}_g$. Thus, for $\theta \in \Theta$, the density of Q_θ with respect to μ is

$$\frac{dQ_\theta}{d\mu}(x) = q_\theta(x) = \exp\{\langle T(x), \theta \rangle - \psi(\theta)\} = Prob\{X = x\}.$$

Let $\mathcal{T} = \{t \in \mathbb{R}^k : t = T(x), x \in \mathcal{G}_g\}$ be the range of $T(\cdot)$ and ν the measure on $\mathcal{T}$ induced by μ, namely

$$\nu(t) = \mu\{x \in \mathcal{G}_g : T(x) = t\} = |\{x \in \mathcal{G}_g : T(x) = t\}|, \quad t \in \mathcal{T}.$$

Then, the distribution of $T(X)$ belongs to the exponential family of distributions on $\mathcal{T}$ with base measure ν, natural parameter space Θ and densities

$$p_\theta(t) = \exp\{\langle t, \theta \rangle - \psi(\theta)\}, \quad \theta \in \Theta.$$

Furtheremore, because of the discreteness of the problem,

$$Prob(T(X) = t) = \int_{\{x \in \mathcal{G}_g : T(x) = t\}} q_\theta(x) d\mu(x) = \sum_{\{x \in \mathcal{G}_g : T(x) = t\}} q_\theta(x) = p_\theta(t)\nu(t).$$

Provided that the network statistics are affinely independent, as it is the case for the examples given above and as it can always be assumed through reduction to minimality, the convex support $P = \text{convhull}(\mathcal{T})$ is a k-dimensional polytope. Furthermore, since ν has finite support, it is easy to see that the assumptions (A1)–(A4) of section 2.1 are verified, and the theory developed above applies.

Despite its simplicity and interpretability, we need to emphasize that ERG modeling based on simple, low dimensional network statistics such as the ones

described above can be rather coarse. In fact, those ERG models are invariant with respect to the relabeling of the nodes and even to changes in the graph topologies, depending on the network statistics themselves. As a result, they do not specify distributions over graphs per se, but rather distributions over large classes of graphs having the same network statistics. Consequently, as elucidated in the example we are about to present (and as we repeatedly observed in our experiments using other network statistics, not reported here), it may very well be the case that many graphs having very different topologies still belong to the same class and, therefore, are considered as equivalent. While this feature may be well suited for defining distributions over large thermodynamic ensembles in statistical physics, its use in other contexts in which the nodes are not interchangeable may be questionable. This is certainly not a common feature of all ERG models: for example, the p_1 model by Holland and Leinhardt (1981) and the Markov graphs by Frank and Strauss (1986) are based on much finer network statistics whose dimension, unlike the aggregate statistics reported above, increases with the size of the network. These more complex models represent explicitly distributions of individual networks rather than of classes on networks: both p_1 and Markov graph models are log-linear models over the probability of edges (see Fienberg and Wasserman, 1981). However, they also present difficulties. In fact, not only is the MLE not likely to exist if the observed network is even moderately sparse, but the asymptotics of these models as g grows remains unknown (see, e.g. Haberman, 1981, for some comments on p_1 models). While the theory developed in the previous sections apply to all ERG models, our analysis below is more directly relevant to models arising from simpler network statistics quantifying macroscopic properties of the network.

3.2. Our running example

We will be using throughout the example of a ERG model on $\mathcal{G}_9$ with two-dimensional network statistic $T(x) = (T_1(x), T_2(x)) \in \mathbb{N}^2$, where $T_1(x)$ is the number of edges and $T_2(x)$ is the number of triangles. Note that this model is not hierarchical in the sense of Bishop et al. (1975) and Lee and Nelder (1996), since we do not include the network statistic for the number of 2-stars, which lie intermediate to edges and triangles. The lack of hierarchical model structure affects the interpretation of the exponential family parameters corresponding to $T(x)$ but turns out not to be the cause of the degeneracies we illustrate. We have actually produced similar results for models which are fully hierarchical, but the results are easier to demonstrate in the context of this ERG model with a two-dimensional network statistic.

The number of distinct graphs for this $\mathcal{G}_9$ example is 2^{36}, while the number of two-dimensional distinct network statistics is only 444. The natural parameter space is $\mathbb{R}^2$. The support of the distribution of $T(X)$ is shown in figure 1. The convex support for the induced family of distributions of network statistics is a polygon with 6 edges, whose boundary is depicted with the red solid line. Out of the possible 444 points, 29 actually lie on the boundary. The fully connected
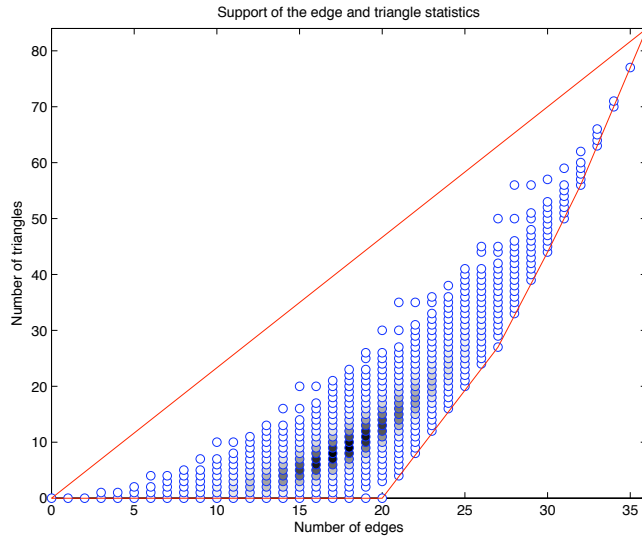
FIG 1. *Support of the distribution of the network statistics for the ERG model on $\mathcal{G}_9$ described in section 3.2. The color shading indicates the square root of the relative frequency of each point, namely $\nu(t)$ (darker colors correspond to higher-frequency values of t). The solid red line is the boundary of the convex support.*

graph has $\binom{9}{2} = 36$ edges and $\binom{9}{3} = 84$ triangles. The induced base measure ν for this family, i.e. the frequencies of each possible pair of network statistics, is indicated by the color shading of the circles. The maximal value of $\nu(t)$ is $1, 876, 664, 161$, the median value is $2, 741, 130$, while the first and third quartiles are $545, 265$ and $79, 674, 084$, respectively. Figure 2 shows a plot of the empirical quantile function for $\nu(t), t \in \mathcal{T}$, which indicates that few network configurations are much more frequent than others.

3.3. Degeneracy

The notion of *degeneracy* is central to ERG modeling, and has been investigated in various forms in the more recent literature. See Snijders (2002), Robins et al. (2007b), Robins et al. (2007a) and, in particular, Handcock (2003) and Hunter et al. (2008), just to mention a few. Degeneracy refers quite broadly to a variety of features, typically undesirable and surprising, of ERG models that have been observed empirically. In the literature, degeneracy (or near degeneracy) is used to describe any of the following, often interrelated, phenomena:

1. when a combination of ERG parameters θ implies that only a small number of graphs have substantial non-zero probabilities; often, these graphs are radically different from each other, for example the empty graph and the fully connected graph;
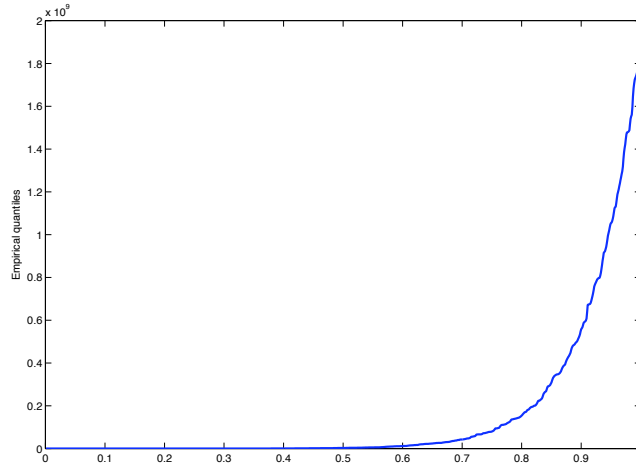
FIG 2. *Empirical quantiles of the values $\{\nu(t), : t \in \mathcal{T}\}$ for the base measure of the family described in section 3.2.*

2. when the MLE of θ does not existent or is hard to obtain, or the MLE of θ as computed by MCMC methods fails to converge or appears to converge extremely slowly;

3. when the estimate of θ would make the observed network very unlikely.

Each of the situations just described offers strong evidence of misspecifcation or, at the very least, of the inability of the model to describe in a realistic fashion the observed network. To our knowledge, Handcock (2003) is the only attempt to characterize degeneracy in a theoretical way, at least the kind of degeneracy yielding unstable maximum likelihood estimates, with emphasis on MCMC-based MLE methods.

3.4. Degeneracy via entropy functions

We base our analysis on a basic observation: a common feature of all the various instances of degenerate ERG models is that the corresponding distributions are highly concentrated on network configurations associated to a small number of network statistics. Therefore, in order to capture the overall degree of concentration of the family $\mathcal{P}$, we turn to Shannon's entropy function, the rationale being that degenerate models have lower entropy.

Shannon's entropy function $S \colon \Theta \to \mathbb{R}$ is defined as

$$S(\theta) = - \sum_{x \in \mathcal{G}_g} q_\theta(x) \log q_\theta(x) = - \sum_{t \in \mathcal{T}} p_\theta(t) \log p_\theta(t) \nu(t),$$

where the second summation involves a much smaller number of terms. Notice that, for every $\theta \in \Theta$,

$$0 \leq S(\theta) \leq \binom{g}{2} \log 2,$$

the lower and upper bounds corresponding to a degenerate distribution with point mass at one graph, and to the uniform distribution over $\mathcal{G}_g$ (which is within the family if $\nu(t)$ is constant across $\mathcal{T}$ and $\theta = 0$), respectively. Furthermore, as ψ is an analytic function of θ, for every $\theta \in \Theta$, $S(\theta)$ is a smooth function of θ.

Noting that $\lim_{x \to 0} x \log x = 0$ and using the fact that $S(\theta)$ is bounded, by the dominated convergence theorem corollary 2.6 yields that, for every $(\theta, \{\rho_n\}, d)$-sequence $\{\theta_n\}$ with $d \in \mathrm{ri}(N_F)$,

$$\lim_n S(\theta_n) = S_F(\theta_F) \equiv - \int_{\mathcal{T}} p_{\theta_F}(t) \log p_{\theta_F}(t) d\nu_F(t), \tag{16}$$

for every face F of P.

On the other hand, because of the correspondence between natural and mean value parameters, the entropy function can be equivalently represented as a function over P. More precisely, we define $V : \mathrm{P} \mapsto \mathbb{R}$ as follows: if $\mu \in \mathrm{ri}(\mathrm{P})$,

$$V(\mu) = S(\theta),$$

where $\mu = \nabla \psi(\theta)$, while, for $\mu_F \in \mathrm{ri}(F)$,

$$V_F(\mu_F) = S_F(\theta_F),$$

where $\mu_F = \nabla \psi^F(\theta_F)$. Thus, if $\{\theta_n\}$ is a $(\theta, \{\rho_n\}, d)$-sequence with $d \in \mathrm{ri}(N_F)$ and if $\mu_n = \mathbb{E}_{\theta_n}(T(X))$, from equation (16) we obtain that

$$\lim_n V(\mu_n) = V_F(\mu_F),$$

where $\mu_F = \lim_n \mu_n$, with $V(\mu)$ a smooth function of μ. Thus, we conclude that $S(\cdot)$ and $V(\cdot)$ have homeomorphic graphs and, therefore, they convey the same information.

Below, we use both entropy functions to illustrate the theory developed in section 2.3 and to provide some characterizations of degeneracy.

We start with figures 3 and 4. The latter displays the entropy function $S(\theta)$ for the ERG model on $\mathcal{G}_9$ with network statistics taking values in $\mathbb{N}^2$, as described in section 3.2, and for values of θ in the rectangle $[10, 25] \times [-25, 10]$. The equivalent entropy function over the mean value space $V(\mu)$ is displayed in figure 3, for the mean value parameters $\{\mu : \mu = \nabla \psi(\theta), \theta \in [10, 25] \times [-25, 10]\}$. Figures 4 and 3 offer two equivalent views of the exponential family $\mathcal{P}$ via the entropy functions $S(\theta)$ and $V(\mu)$. The mean value view in figure 3 is straightforward to interpret: the entropy function is a well behaved, strictly concave function that changes smoothly as the mean parameter varies inside the relative interior of P. Distributions with mean value parameters lying well inside the cloud of points describing the support of the family have higher entropy, as
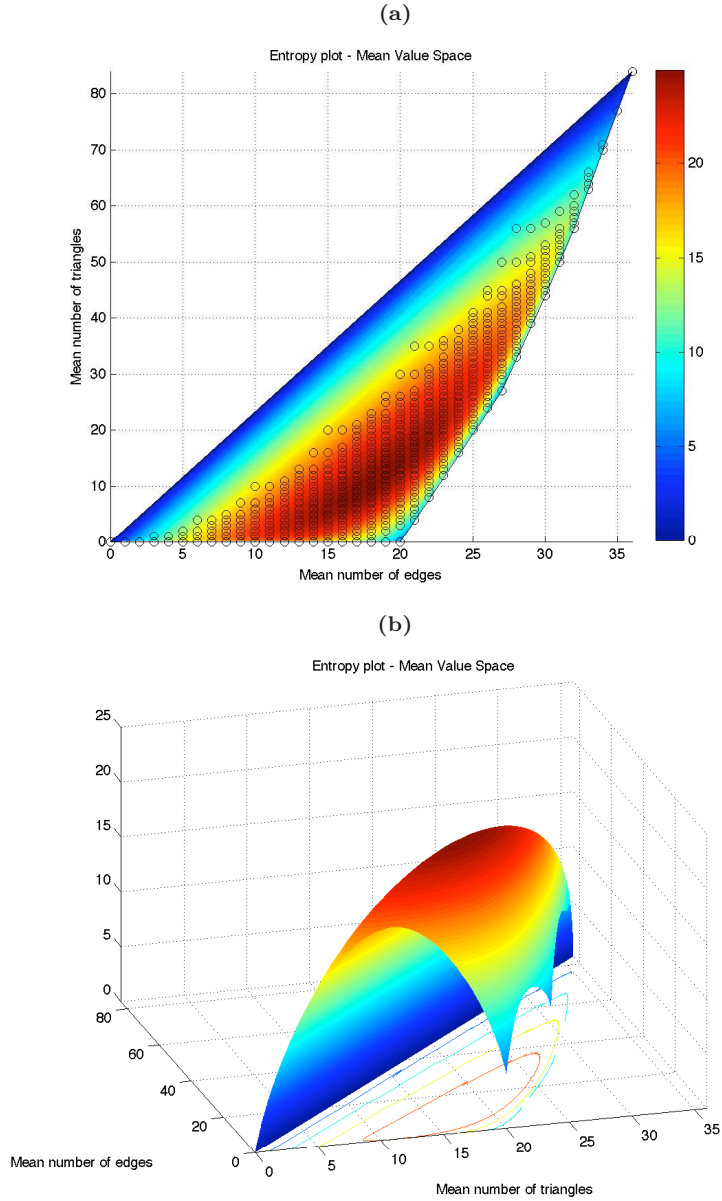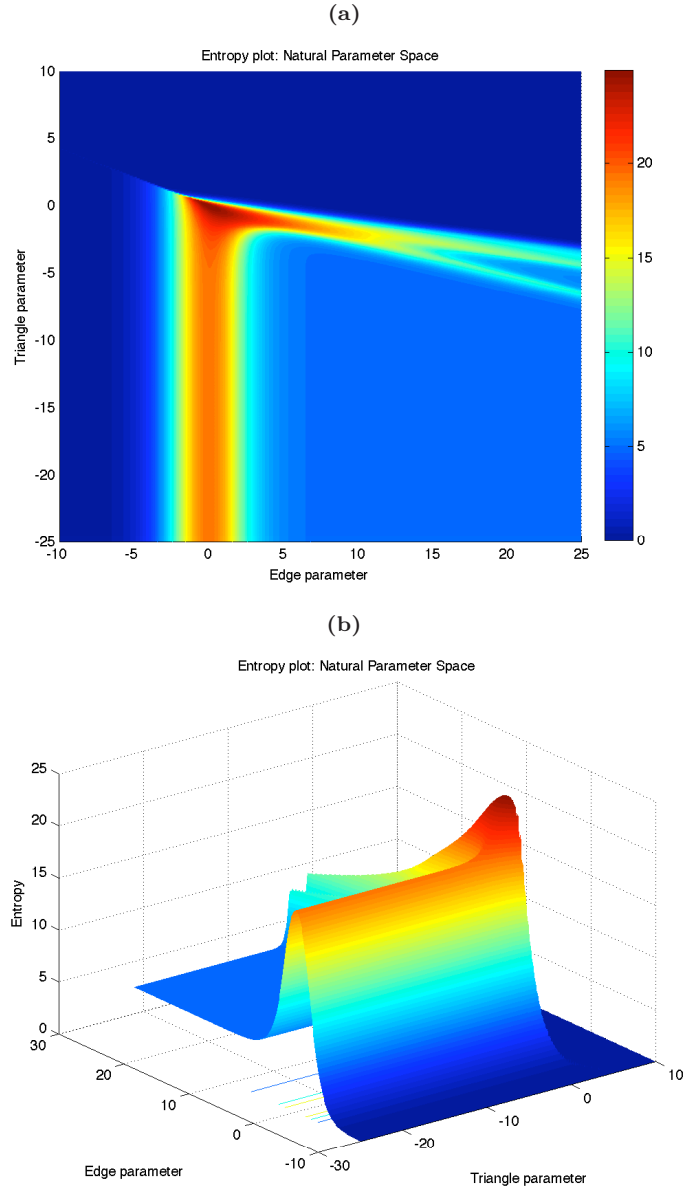
(a)

(b)

FIG 3. *Plots of the entropy function $V(\cdot)$ under mean value parametrization for the ERG model of section 3.2. Part* **a)**: *2-dimensional plot over the convex support* P; *the points correspond to the support of the family. Part* **b)**: *surface plot.*

their mass is distributed across a larger number of network configurations. In contrast, distributions with mean value parameters that are far removed from that cloud, including points very close to or on the boundary of P, have lower

(a)

(b)

FIG 4. *Plots of the entropy function $S(\cdot)$ under natural parametrization for the ERG model of section 3.2. Part **a**): 2-dimensional plot over a square of the natural parameter space. Part **b**): surface plot.*

entropy. It is worth pointing out that, for this specific family, the points in the support are closer to the lower boundary of the polygon P, while the side of P determined by the convex hull of points corresponding to the empty and com-

plete graph is significantly distant from the support. This phenomenon becomes more pronounced as g grows, so that this family will include many distributions, whose mean value parameters belong to a region far removed from the support, that would not provide a satisfactory or realistic explanation of any observed network, a feature that is often associated with degeneracy.

In striking contrast, the natural parameter view of figure 4 does not lend itself to immediate interpretations. In fact, although $S(\theta)$ and $V(\mu)$ are smooth functions related via the homeomorphism (5), $S(\theta)$ displays drastic localized behaviors, including multiple local maxima. In particular, the function $S(\theta)$ exhibits sharp changes and high-peaked ridges shooting at infinity along which it remains roughly constant. Furthermore, small variations in the natural parameter values cause big changes in the values of the entropy function, thus making this ERG model rather unstable, in the sense that neighboring parameters specify very different distributions, or at least distributions with different entropies. These features may in fact fall under the general umbrella of degeneracy, as described in section 3.3. Finally, we remark that the portion of the natural parameter space containing parameter points that produce more realistic distributions with higher entropy values is relatively small, a characteristic that emerged from the inspection of figure 3 as well. In addition, the entropy function remains relatively high along some rays leaving the origin and shooting to infinity. We remark that figure 4 matches quite closely analogous plots, not based on Shannon's entropy, for the same ERG model on graphs with 7 nodes by Handcock (2003), although the interpretation of the plots using normal cones, as described below, is missing.

Figure 5 shows all the possible MLEs corresponding to the 415 points in the support of $\mathcal{E}_P$ that are in he interior of P. These points are all the estimates that can be obtained by the maximum likelihood procedure, so that, although the family $\mathcal{E}_P$ contains many other distributions, inference is only restricted to the 415 distributions identified by the MLEs, whose entropies are displayed in the figure.

Part of the seemingly strange behavior of $S(\theta)$ can however be explained using the results derived in the previous section. To that end, the convex support of figure 1, can be represented in $\mathcal{V}$ form either as the convex hull of its vertices, namely

$$\mathrm{P} = \operatorname{convhull}\{(0,0),(20,0),(27,27),(30,44),(32,56),(36,84)\}$$

or, equivalently, using the $\mathcal{H}$-representation, as the solution set of a system of linear inequalities, i.e.

$$\mathrm{P} = \{t \in \mathbb{R}^2 \colon \mathrm{A}t \leq b\},$$

where

$$\mathrm{A} = \begin{bmatrix} 0 & -1 \\ 27 & -7 \\ 17 & -3 \\ 6 & -1 \\ 7 & -1 \\ -21 & 9 \end{bmatrix} \quad \text{and} \quad b = \begin{bmatrix} 0 \\ 540 \\ 432 \\ 136 \\ 168 \\ 0 \end{bmatrix}.$$
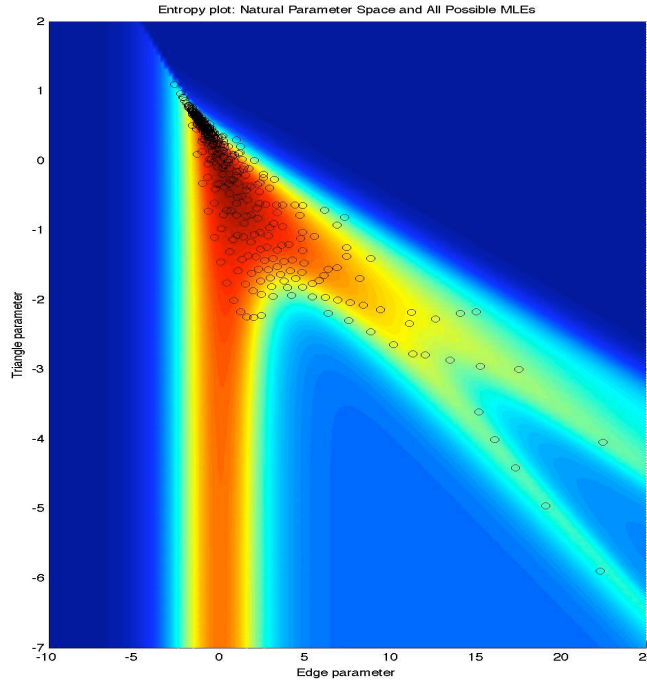
FIG 5. *All possible MLEs of the natural parameters for the ERG model of section 3.2 superimposed over the entropy plot of $S(\cdot)$.*

The rows of A identify the outer normals to the 6 sides of the polygon P and generate the normal cones to the edges of P. The normal cone of a vertex of P is the conic hull of the outer normals to the edges incident to that vertex. For example, the normal cone of the vertex $(0,0)$ is

$$\text{cone}\left\{(0,-1),(-21,9)\right\}$$

The convex support P and its outer normals are shown in figure 6. The normal fan of P, i.e. the collections of all the cones with apex at 0 identified by the outer normals of P, is displayed in figure 7. It is immediate to see that the normal fan induces a partition of $\mathbb{R}^2$ in 6 regions, which are all polyhedral cones. Every region corresponds to one vertex of P. The boundaries between adjacent regions are rays emanating from the origin (thus, they are also polyhedral cones); every such ray corresponds to one edge of P.

Figure 8 shows the entropy plot over the subset $[10,25] \times [-25,10]$ of the natural parameter space with, superimposed, the normal fan of P, which we centered at the origin, the point of maximal entropy. As prescribed by corollary 2.6, the outer normals to P are precisely the directions along which the closure
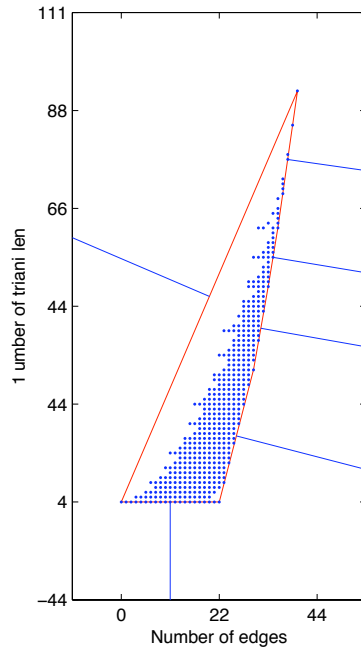
FIG 6. *Convex support and its outer normals for the ERG model of section 3.2.*
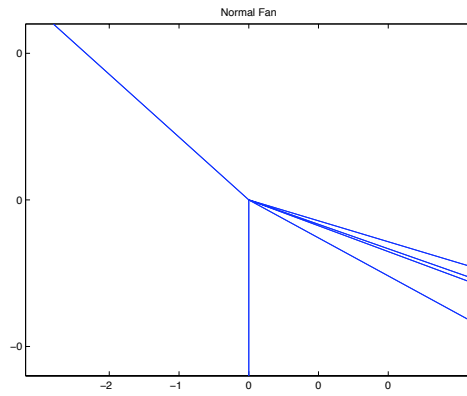

FIG 7. *Normal fan of the convex support for the ERG model of section 3.2. For this particular example, the normal fan is constructed by collecting all the outer normals of* P *shown in figure 6 and displaying them as rays leaving the origin. Note that the scales of the axes are set differently.*
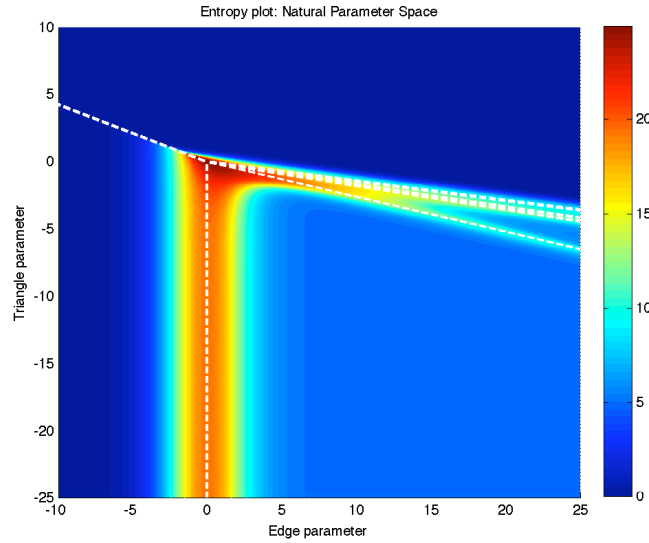
FIG 8. *Entropy plot of $S(\cdot)$ with, superimposed, the normal fan of* P *for the ERG model of section 3.2.*

of the original family $\mathcal{E}_{\mathrm{P}}$ is realized, by adding the families $\mathcal{E}_F$, as F ranges over the proper faces (in this case, edges and vertices) of P. These directions, starting at the origin, match perfectly the ridges of figure 4, along which the entropy function seems to converge to some fixed value. This is the case because any sequence $\{\theta_n\}$ along the outer normal of some edge F will eventually no longer identifies distributions from the original family $\mathcal{E}_{\mathrm{P}}$, but just *one* distribution in $\mathcal{E}_F$ supported on F. Consequently, the entropy function does not change because, for all n large enough, θ_n specifies almost the same distribution.

Figures 9, 10 and 11 offer other two pictorial representations of corollary 2.6. These plots were obtained using the MATLAB GUI available at http://www.stat.cmu.edu/~arinaldo/ERG/ (see section 4 below). The left side of each plot shows the entropy function for the family of section 3.2 along with the outer normals of P leaving the original. The white circles represent the selected natural parameter. The plots on the right show the support of the family. The red stars indicate the mean parameter values corresponding to the natural parameters indicated by the white circles on the left side of the figure. Points with darker shaded colors correspond to network statistics receiving high probability under the selected natural parameter.

Part (**a**) of figure 9 shows a distribution with high entropy, corresponding to a mean value parameter in the relative interior of P. In contrast, in parts (**b**), (**c**) and (**d**) the natural parameter is selected as d, with d a point in the relative interior of the 2-dimensional normal cone of the vertex of coordinates $(0, 0)$, which identifies the empty graph. Consequently, the entropy is almost 0,

as the associated distribution will put almost all its mass on that vertex of P. Notice that, even though the selected natural parameters from part (**b)**, (**c)** and (**d)** are very different from each other, because they are far away from the set of parameters producing nondegenerate distributions and because they all to lie inside the normal cone of the vertex $(0, 0)$, they parametrize essentially the same degenerate distribution on the empty graph.

Figure 10 part (**a)** shows the same phenomenon, but for the different degenerate distribution putting virtually all its mass on the complete graph, which corresponds to the vertex $(36, 84)$. As with figure 9, notice that the natural parameter is a point inside the normal cone of that vertex and essentially any point in the upper triangular blue part of the entropy plot (which is, effectively, the relative interior of the associated normal cone) would parametrize this distribution. Part s(**b)** and (**c)** show other degenerate distributions over the vertex of P identified by points inside the interiors of the corresponding normal cones. Figure 11 instead displays similar plots for a selection of natural parameters corresponding to directions lying on the normal cones, i.e. the outer normals, of some of the edges of P.

4. Discussion

The purpose of this article is two-fold. First, for the class of discrete linear exponential families with polyhedral convex support, we have characterized the extended family using the normal fan to the convex support. While deep and exhaustive results about closures of general exponential families exist in the literature, our restriction to families with polyhedral support allowed us to obtain a more refined and explicitly geometrical description. In particular, our findings allowed us to gain a better understanding of the geometric and statistical properties of these families, as well as on the theoretical and algorithmic aspects of computing extended maximum likelihood estimates. v Our second goal was to study the behavior and statistical properties of ERG models, that have seen widespread use in the statistical analysis of social networks data. To that end, we applied the theoretical results derived in the first part of the article to one ERG model on the set of graphs with 9 nodes. Despite our analysis being mostly graphical (due to the lack of a closed-form expression for the log-partition function), it captures a few interesting features of this model, some of which accounts for the seemingly strange behaviors that ERG models have been known to exhibit in practice, and generically termed degeneracy. Our investigation indicated that this type of behavior is, in fact, not unusual, and can be explained by the properties of linear discrete exponential families.

Although by no means do we mean to imply that all ERG models exhibit the same type of features we have indicated for the one example we study in this article, nonetheless we believe our conclusions are not just specific to this model and could apply more widely to ERG models in which the nodes are treated as interchangeable and whose individual characteristics are not taken into account. This position seems to be corroborated by some of the empirical and theoretical
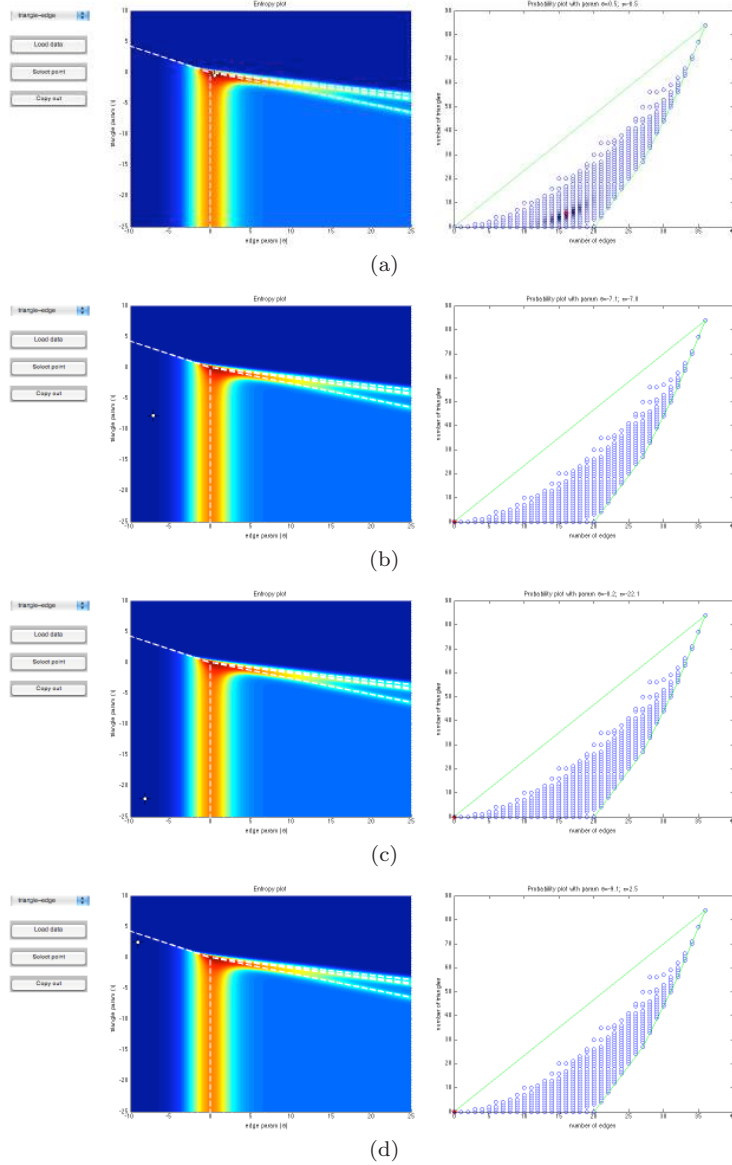
(a)

(b)

(c)

(d)

Fig 9. *Various distributions parametrized by points in the natural parameter space for the ERG model of section 3.2. The plots on the left are the entropy plots; the white points indicate the selected distributions. The plots on the right all display convex support. The red crosses represent the mean value parameters corresponding to the selected natural parameters, while the darker shading indicates network statistics configurations that are very probably under the selected parameters. Part (a): distribution with high-entropy with mean value parameter inside* P. *Parts (b), (c) and (d): natural parameters all specifying distributions with virtually all of the total mass on the empty graph.*
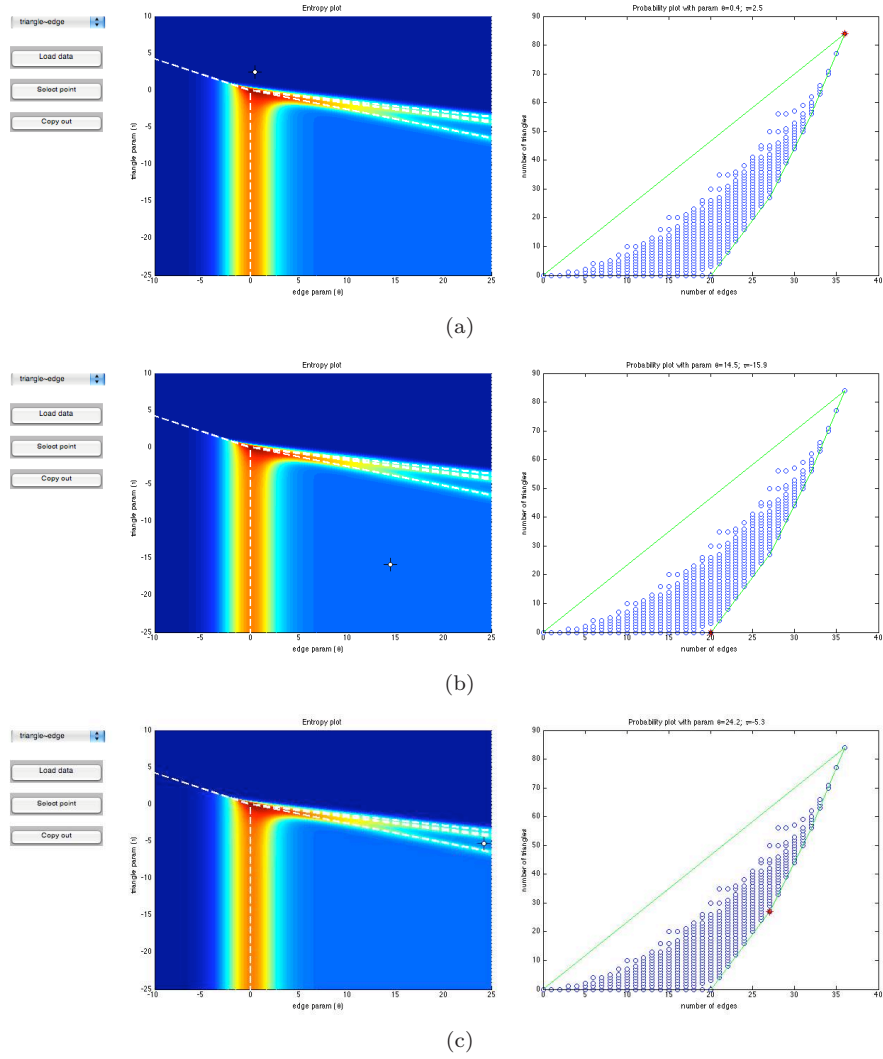
FIG 10. *Three degenerate distributions over three vertices of* P. *See the caption of figure 9.*

analyses of ERG models existing in the social network literature we have cited, as well as by similar experimentations we conducted with other ERG models (for example using k-stars statistics). On a related note, the application presented here are particularly relevant to ERG models built around network statistics that describes macroscopic features of the networks and whose dimension does not grow with the number of nodes. However, our results are relevant to more complex models, such as the p_1 model of Holland and Leinhardt (1981), which has node-specific parameters and whose likelihood is based on an assumption
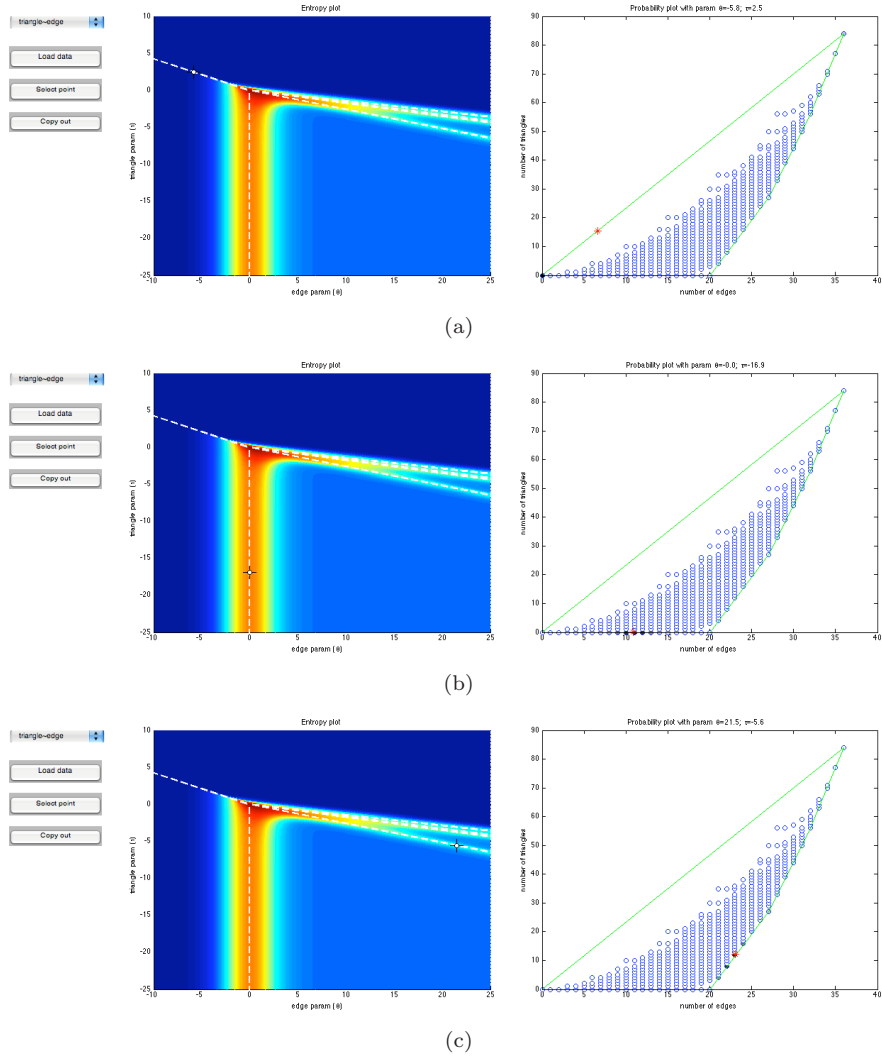
(a)

(b)

(c)

Fig 11. *Three degenerate distributions supported over three different edges of of* P. *See the caption of figure 9.*

of dyadic independence. For these models with many parameters, degeneracy is typically due to nonexistence of the MLE, which is very likely to occur if the network is even mildly sparse.

Appendix A: Proofs

Proof of Lemma 2.2. Let $\theta_1 \in \Theta_F$ and $\zeta \in \text{lin}(N_F)$ and consider the point $\theta_2 = \theta_1 + \zeta$. We first show that $\theta_2 \in \Theta_F$ and $P_{\theta_1} = P_{\theta_2}$. Because $\zeta \in \text{lin}(N_F)$,

there exist some scalars $c_1 \ldots, c_{m_F}$ such that

$$\zeta = \sum_{i=1}^{m_F} c_i a_i.$$

Therefore, almost everywhere ν_F,

$$\langle \zeta, x \rangle = \sum_{i=1}^{m_F} c_i b_i \equiv C.$$

Then,

$$\psi_F(\theta_2) = \log \int_F \exp^{\langle \theta_1 + \zeta, x \rangle} d\nu_F(x) = \log \int_F \exp^{\langle \theta_1, x \rangle} d\nu_F(x) + C = \psi_F(\theta_1) + C.$$

As both $\psi_F(\theta_1)$ and C are finite, it follows that $\psi_F(\theta_2) < \infty$ and, therefore, $\theta_2 \in \Theta_F$. It is now easy to conclude that $P_{\theta_1} = P_{\theta_2}$ because, almost everywhere ν_F,

$$p_{\theta_2}(x) = \exp^{\langle \theta_1 + \zeta, x \rangle - \psi_F(\theta_2)} = \exp^{\langle \theta_1, x \rangle + C - \psi_F(\theta_1) - C} = p_{\theta_1}(x).$$

We now show that if $P_{\theta_1} = P_{\theta_2}$ and $\theta_1 \neq \theta_2$, then $\theta_1 - \theta_2 \in \lin(N_F)$. By Radon-Nykodin theorem, this occurs if and only if

$$\langle x, \theta_1 - \theta_2 \rangle = \psi^F(\theta_1) - \psi^F(\theta_2) = D$$

for some constant D, almost everywhere ν_F. As ν_F has support contained in F and F is defined by (2), the previous equality is equivalent to $\theta_1 - \theta_2 \in \lin(N_F)$, thus completing the proof of the lemma.

As for (9), since P is full-dimensional and, almost everywhere ν_F, $A_F x = b_F$, we have, for any $\theta \in \Theta_F$,

$$0 = \mathrm{Var}_\theta\left(\langle a, X \rangle\right) = a^\top I_F(\theta) a$$

if and only if $a = \lin(N_F)$. This implies that $\mathrm{rank}(I_F(\theta)) = \dim\left(\lin(N_F)^\perp\right) = \dim(F)$. ∎

Proof of Lemma 2.3. Arguing by contradiction, suppose that, for all n large enough, θ_n belongs to a compact, hence bounded, set C. The facts that $\nabla \psi(\theta) = \mathbb{E}_\theta[X] \in \mathrm{ri}(P)$ Θ, for each $\theta \in \Theta$ with finite norm, and that $\mathrm{ri}(P)$ and Θ are homeomorphic, imply that $\{\nabla\psi(\theta), : \theta \in C\}$ is a compact subset of $\mathrm{ri}(P)$. Then, because $\|\nabla\psi(\theta) - \mu_F\|_2$ is a continuous function of θ, for all $\theta \in \Theta$, $\inf_{\theta_n \in C} \|\nabla\psi(\theta_n) - \mu_F\|_2 = \|\nabla\psi(\theta^*) - \mu_F\|_2$ for some $\theta^* \in C$. But then, $\nabla\psi(\theta^*) \equiv \mu^* \in \mathrm{ri}(P)$ so that, $\|\mu^* - \mu_F\|_2 > \epsilon > 0$ for some ϵ, which produces a contradiction. ∎

Proof of theorem 2.4. Throughout the proof, we will write $S_{k-1} = \{x \in \mathbb{R}^k : \|x\|_2 = 1\}$.

In the proof we will make use repeatedly of the following decomposition. For any point $x_0 \in \mathrm{P}$ and proper face F of P, we will write

$$p_{\theta_n}(x_0) = \frac{\exp^{\langle \eta, x_0 \rangle}}{A_{0,n}(x_0, F) + A_{>,n}(x_0, F) + A_{<,n}(x_0, F)}, \tag{17}$$

where

$$A_{0,n}(x_0, F) = \int_{\{x \colon A_F(x - x_0) = 0\}} \exp^{\langle \eta, x \rangle + \rho_n \langle d_n, x - x_0 \rangle} \, d\nu(x),$$

$$A_{>,n}(x_0, F) = \int_{\{x \colon A_F(x - x_0) > 0\}} \exp^{\langle \eta, x \rangle + \rho_n \langle d_n, x - x_0 \rangle} \, d\nu(x),$$

and

$$A_{<,n}(x_0, F) = \int_{\{x \colon A_F(x - x_0) < 0\}} \exp^{\langle \eta, x \rangle + \rho_n \langle d_n, x - x_0 \rangle} \, d\nu(x).$$

Notice that, for all n, if $x_0 \in F$, then $A_{>,n}(x_0, F) = 0$, since $\nu\{x \colon A_F(x - x_0) > 0\} = 0$. We will also use the following fact, which stems directly from lemma 2.2: $\exp^{\langle \eta, x \rangle - \psi^F(\eta)} = \exp^{\langle \theta, x \rangle - \psi^F(\theta)} = p_{\theta_F}^F(x)$, almost everywhere ν_F.

1. Part 1.
   We will begin by showing sufficiency. First, we consider the case of a generic point $x_0 \in F$. If, $d_n \in \mathrm{ri}(N_F)$, then, by part *1.* of lemma B2, $\langle d_n, x - x_0 \rangle = 0$ for all $x \in F$, which implies that

   $$A_{0,n}(x_0, F) = \int_F \exp^{\langle \eta, x \rangle} \, d\nu(x) = \exp^{\psi^F(\eta)},$$

   for all n. On the other hand, for any $x \notin F$, since R is a compact subset of $\mathrm{ri}(N_F) \cap S_{k-1}$ and $\{d_n\} \in R$, we have

   $$\sup_n \langle d_n, x - x_0 \rangle \leq \sup_{d \in R} \langle d, x - x_0 \rangle = \langle d_x^*, x - x_0 \rangle,$$

   for some $d_x^* \in R$, which may depend on x. Furthermore, by part *2.* of lemma B2, $\langle d_x^*, x - x_0 \rangle < 0$. Thus, $\rho_n \langle d_n, x - x_0 \rangle \to -\infty$, for each $x \notin F$. for each $x \in \{x \colon A_F(x - x_0) < 0\}$,

   $$\exp^{\langle \eta, x \rangle + \rho_n \langle d_n, x - x_0 \rangle} \leq \exp^{\langle \eta, x \rangle},$$

   whereby

   $$\int_{\{x \colon A_F(x - x_0) < 0\}} \exp^{\langle \eta, x \rangle} \, d\nu(x) \leq \int_{\mathbb{R}^k} \exp^{\langle \eta, x \rangle} \, d\nu(x) = \exp^{\psi(\eta)} < \infty.$$

   Then, by the dominated convergence theorem, we obtain

   $$A_{<,n}(x_0, F) \searrow 0.$$

   Therefore,

   $$A_{0,n}(x_0, F) + A_{<,n}(x_0, F) \searrow \exp^{\psi^F(\eta)},$$

which implies that

$$\lim p_{\theta_n}(x_0) \nearrow \exp^{\langle \eta, x_0 \rangle - \psi^F(\eta)} = p_{\theta_F}(x_0). \tag{18}$$

Next, let $x_0 \in \mathrm{P} \cap F^c$ and notice that

$$A_{>,0}(x_0, F) + A_{0,n}(x_0, F) + A_{<,n}(x_0, F)$$
$$\geq A_{>,n}(x_0, F) \geq \int_F \exp^{\langle \eta, x \rangle + \rho_n \langle d_n, x - x_0 \rangle} \, d\nu(x),$$

since $F \subseteq \{x \colon A_F(x - x_0) > 0\}$. For any $x \in F$, since $\{d_n\} \in R$ and R is a compact subset of $\mathrm{ri}(N_F) \cap S_{k-1}$, we get

$$\inf_n \langle d_n, x - x_0 \rangle \geq \inf_{d \in R} \langle d, x - x_0 \rangle = \langle d_x^*, x - x_0 \rangle,$$

for some $d_x^* \in R$, which may depend on x. By lemma B2, part 2., $\rho_n \langle d_x^*, x - x_0 \rangle \to \infty$, for all $x \in F$. But then, as $\nu(F) > 0$ by assumption (A3), we obtain

$$\int_F \exp^{\langle \eta, x \rangle + \rho_n \langle d_n, x - x_0 \rangle} \, d\nu(x) \to \infty,$$

by the monotone convergence theorem. Thus,

$$A_{>,0}(x_0, F) + A_{0,n}(x_0, F) + A_{<,n}(x_0, F) \to \infty, \tag{19}$$

and, therefore, $p_{\theta_n}(x_0) \to 0 = p_{\theta_F}^F(x_0)$.

2. Part 2.
   Suppose that, $\{d_n\} \subset R$, where R is a compact subset of N_F^c. Then, there exists a subsequence $\{d_{n_k}\} \subset \{d_n\}$ such that, for all k large enough, d_{n_k} belongs to a compact set R^* such that either $R^* \subset \mathrm{ri}(N_{F'})$, for some $F' \neq F$, or $R^* \subset (\mathcal{N}(\mathrm{P}))^c$. In the latter case, by part 3., proven below, the numbers $\|\mu_{n_k} - \mu_F\|_2$ grow unbounded and, therefore, (11) is violated. In the former case, by part 1. of the proof, (11) is verified for F', so it cannot be simultaneously verified for F as well. Indeed, p_{θ_n} cannot converge pointwise to both $p_{\theta_F}^F$ and $p_{\theta_{F'}}^{F'}$, which identify different probability distributions with different supports.

3. Part 3.
   We will show that, if $\{d_n\} \subset R$ for some compact subset of $(\mathcal{N}(\mathrm{P}))^c$, then,

$$p_{\theta_n}(x_0) \to 0, \quad \forall x_0 \in \mathrm{P}. \tag{20}$$

This implies that $\|\mu_n\|_2 \to \infty$. Let $x_0 \in \mathrm{P}$. As P is full-dimensional and $d_n \notin \mathcal{N}(\mathrm{P})$, by lemma B2, part 3., the set $S_n = \{x \in \mathrm{P} \colon \langle d_n, x - x_0 \rangle > 0\}$ is non-empty, for each n. Furthermore, since, by assumption,

$$\inf_{d \in R} \inf_{d' \in \mathcal{N}(\mathrm{P})} \|d - d'\|_2 > 0,$$

the set $S = \liminf_n S_n$ is non-empty as well. We now claim that $\nu(S) > 0$. In fact, arguing by contradiction, suppose that $\nu(S) = 0$. Then, there

exists a subsequence $\{d_{n_k}\} \subset \{d_n\}$ such that no point $x \in \text{supp}(\nu)$ can satisfy $\lim_k \langle d_{n_k}, x - x_0 \rangle > 0$. However, since, by assumption (A3), $P = \text{convhull}(\text{supp}(\nu))$, this implies that $P \subseteq \{y \colon \lim_k \langle d_{n_k}, y - x_0 \rangle \leq 0\}$, which in turn implies that $\lim_k d_{n_k} \in \mathcal{N}(P)$, violating the condition that $\{d_n\}$ is bounded away from $\mathcal{N}(P)$. Thus, $\nu(S) > 0$, from which we can conclude that $\lim \inf_n \nu(S_n) \geq \nu(S) > 0$. Then, by the monotone convergence theorem,

$$\int_{S_n} \exp^{\langle \eta, x \rangle + \rho_n \langle d_n, x - x_0 \rangle} \, d\nu(x) \to \infty,$$

Therefore,

$$p_{\theta_n}(x_0) = \frac{\exp^{\langle \eta, x \rangle}}{\int_{S_n} \exp^{\langle \eta, x \rangle + \rho_n \langle d_n, x - x_0 \rangle} \, d\nu(x) + \int_{S_n^c} \exp^{\langle \eta, x \rangle + \rho_n \langle d_n, x - x_0 \rangle} \, d\nu(x)} \to 0,$$

as claimed.

$\square$

Proof of Corollary 2.6. Any direction $d \in \mathbb{R}^k$ is either in $\mathcal{N}(P)$, in which case, it must belong to $\text{ri}(N_F)$ for one face F of P or in $(\mathcal{N}(P))^c$. The results then follow directly from theorem 2.4. $\square$

Proof of Corollary 2.8. If $x \in \text{ri}(P)$, then the MLE exists, is unique and is given by the vector $\widehat{\theta} \in \Theta$ such that $\nabla \psi(\widehat{\theta}) = x$. Equivalently, since in this case $N_P = \{0\}$, invoking corollary 2.6, part *1.*, $-\ell_x$ has no direction of recession. Thus consider the case of $x \in \text{rb}(P)$ and let F be the unique face such that $x \in \text{ri}(F)$. If $d \in \text{ri}(N_F)$, then by corollary 2.6 part *1.*,

$$\lim_{\rho \to \infty} p_{\theta + \rho d}(x) > 0, \tag{21}$$

so (13) holds. Suppose now that $d \in \text{rb}(N_F)$. Let $\mathcal{F}_x = \{F' \colon x \in \text{rb}(F')\}$, with F' being a face of P. By lemma B2, part *4.*,

$$\biguplus_{F' \in \mathcal{F}_x} \text{ri}(N_{F'}) = \text{rb}(N_F),$$

so, if $d \in \text{rb}(N_F)$, then $d \in \text{ri}(N_{F'})$, for some $F' \in \mathcal{F}_x$. By corollary 2.6, part *1.*, almost everywhere $\nu_{F'}$,

$$\lim_{\rho \to \infty} p_{\theta + \rho d} = p_{\theta F'}^{F'}.$$

Since $x \in F'$, we have $\nu_{F'}(x) > 0$, which implies, $p_{\theta F'}^{F'}(x) > 0$ and, consequently, (21). Thus d is also a direction of recession and we have shown that any point in N_F is a direction of recession for $-\ell_x$,

It remains to be shown that equation (13) is not verified if $d \notin N_F$. If $d \notin \mathcal{N}(P)$ corollary 2.6 part *2.* yields

$$\lim_{\rho \to \infty} p_{\theta + \rho d}(x) = 0,$$

hence $-\ell_x(\theta) \to \infty$, so d is not a direction of recession. If instead $d \in \mathcal{N}(\mathrm{P}) \cap N_F^c$, then it must be the case that $d \in \mathrm{ri}(N_F^*)$, for some face F^* such that $F \cap F^* = \emptyset$, otherwise $N_{F^*} \subset \mathrm{rb}(N_F)$ (see, e.g., lemma B2, part *4.*). Thus, by corollary 2.6, part *1.*,

$$\lim_{\rho \to \infty} p_{\theta + \rho d}(x) = p_{\theta_{F^*}}^{F^*}(x) = 0,$$

because $x \notin F^*$, while $p_{\theta_{F^*}}^{F^*}(x) > 0$ only if $x \in F^*$. As a result, (21) does not hold, so that d does not satisfy (13) and is not a direction of recession. $\square$

Proof of Corollary 2.9. The only interesting case is when $x \in \mathrm{ri}(F)$, for some proper face F, otherwise $\mathcal{N}(\mathrm{P}) = \{0\}$, and $-\ell_x$ has no directions of recession, as the MLE exists. For every $\theta \in \Theta$, let $\{\theta_n\}$ be a $(\theta, \{\rho_n\}, d)$-sequence. By corollary 2.8, we need to consider only the case $d \in N_F$. If $d \in \mathrm{ri}(N_F)$, by lemma B1, $p_{\theta_n}(x) \nearrow p_{\theta_F}^F(x)$. Now suppose that $d \in \mathrm{rb}(N_F)$. Then, $d \in \mathrm{ri}(N_{F^*})$ for some face F^* such that $F \subset F^*$. Another application of lemma B1, yields $p_{\theta_n}(x) \nearrow p_{\theta_{F^*}}^{F^*}(x)$. However,

$$p_{\theta_{F^*}}^{F^*}(x) = \exp^{\langle \theta, x \rangle - \psi^{F^*}(\theta)} < \exp^{\langle \theta, x \rangle - \psi^F(\theta)} = p_{\theta_F}^F(x),$$

since

$$\exp^{\psi^{F^*}(\theta)} = \int_{F^*} \exp^{\langle \theta, z \rangle} d\nu(z) \geq \int_F \exp^{\langle \theta, z \rangle} d\nu(z) = \exp^{\psi^F(\theta)}.$$

Thus, $\sup_{\theta \in \Theta} p_\theta(x) = p_{\gamma_F}^F(x)$ for some $\gamma^F \in \Theta_F$. But $\sup_{\gamma_F \in \Theta_F} p_{\gamma_F}(x) = p_{\widehat{\theta}_F}^F(x)$, since only the points $\theta \in \widehat{\theta}_F$ satisfy the first order optimality conditions (6). The result follows. $\square$

Proof of Corollary 2.10. Parts *i)* and *ii)* follow from lemma 2.2 and results of Csiszár and Matúš (2003, 2005). Part *iii)* is a direct consequence of part *i)*. $\square$

Proof of Corollary 2.11. For any $\theta \in \Theta$, the (i, j)-th entry of $I(\theta)$ is (see, e.g., corollary 2.3 in Brown, 1986)

$$I_{i,j}(\theta) = \frac{\partial}{\partial \theta_i \partial \theta_j} \psi(\theta).$$

From the proof of theorem 2.4, $\psi(\theta_n) \to \psi^F(\theta + \zeta)$, for every $\zeta \in \mathrm{lin}(N_F)$. Then, by the analytic properties of the cumulant generating function (see, e.g. Brown, 1986, Chapter 2), we obtain

$$\lim_n \frac{\partial}{\partial \theta_i \partial \theta_j} \psi(\theta_n) = \frac{\partial}{\partial \theta_i \partial \theta_j} \lim_n \psi(\theta_n) = \frac{\partial}{\partial \theta_i \partial \theta_j} \psi^F(\theta + \zeta) = I_F(\theta + \zeta),$$

for every $\zeta \in \mathrm{lin}(N_F)$, hence the statement is proved. $\square$

Appendix B: Additional proofs

We need the following lemma in the proof of corollary 2.9

Lemma B1. Under the conditions of corollary 2.6, $p_{\theta_n} \nearrow p_{\theta_F}^F$, a.e. ν_F, if and only if $d \in \mathrm{ri}(N_F)$.

Proof. The claim follows from equation (18) in the proof of theorem 2.4, which holds for all $x \in F$, thus almost everywhere ν_F. □

Below, we collect some basic facts about the normal fan and normal cones needed in our proofs. We will say that a vector d is normal to the hyperplane H if $\langle d, x - y \rangle = 0$ for all $x, y \in H$.

Lemma B2. Let P be full-dimensional and let F be a face of P.

1. For any $x_0 \in F$, $\langle a^F, x - x_0 \rangle = 0$ for all $x \in F$ and $\langle a^F, x - x_0 \rangle < 0$ for all $x \notin F$ if and only if $a^F \in \mathrm{ri}(N_F)$.
2. For any $x_0 \notin F$, $\langle a^F, x - x_0 \rangle > 0$ for all $x \in F$ and $\langle a^F, x - x_0 \rangle \leq 0$ for all $x \notin F$ if and only if $a^F \in \mathrm{ri}(N_F)$.
3. If $d \notin \mathcal{N}(P)$, then, for any $x_0 \in P$,

$$P = S_{>,x_0} \uplus S_{=,x_0} \uplus S_{<,x_0}$$

    where $S_{>,x_0}$, $S_{=,x_0}$ and $S_{<,x_0}$ are disjoint, non-empty sets given by $\{x \in P \colon \langle d, x - x_0 \rangle > 0\}$, $\{x \in P \colon \langle d, x - x_0 \rangle = 0\}$ and $\{x \in P \colon \langle d, x - x_0 \rangle > 0\}$, respectively.
4. $\mathrm{rb}(N_F) = \biguplus_{F' \colon F' \supset F} \mathrm{ri}(N_{F'})$, where the disjoint union ranges over all the faces F' of P.
5. $N_F = \mathrm{cone}\,(a_1, \ldots, a_{m_F})$, where a_i denotes the transpose of the i-th row of the submatrix A_F given in (2), $i = 1 \ldots, m_F$.

Proof. Recall that, since P is full-dimensional, there is no vector $d \neq 0$ such that $\langle d, x - x_0 \rangle = 0$ for all pairs $x, x_0 \in P$.

1. First we show sufficiency. If $a^F \in \mathrm{ri}(N_F)$, then a^F is a conic combination of all the rows of A_F with positive coefficients. Therefore, $\langle a^F, x - x_0 \rangle = 0$ for all $x \in F$, by the definition of F, and $\langle a^F, x - x_0 \rangle < 0$ for all $x \notin F$, since, in this case, $\langle a, x - x_0 \rangle < 0$ for some row a of A_F. As for necessity, if $a^F \in N_F$, then $\langle a^F, x - x_0 \rangle < 0$ for all $x \in \mathrm{ri}(P)$. However, if $a^F \in \mathrm{rb}(N_F)$, then $\langle a^F, x - x_0 \rangle = 0$ for all $x \in F'$, where F' is the face of P such that $a^F \in \mathrm{ri}(N_{F'})$. But then, since $F \subset F'$, there exists a $x \notin F$ for which $\langle a^F, x - x_0 \rangle = 0$, which would produce a contradiction. Thus $a^F \notin \mathrm{rb}(N_F)$.
2. The proof is analogous to the previous case and is omitted.
3. Since d is not normal to any supporting hyperplane, the hyperplane $H = \{x \colon \langle d, x - x_0 \rangle = 0\}$ intersects P is in its relative interior, and P must have non-empty intersections with both the halfspaces $\{x \in \mathbb{R}^k \colon \langle d, x - x_0 \rangle > 0\}$ and $\{x \in \mathbb{R}^k \colon \langle d, x - x_0 \rangle < 0\}$ cut out by H.
4. The claim follows directly from the definition of N_F and the fact that $\mathcal{N}(P)$ is a polyhedral complex (see, e.g., Sturmfels, 1995), thus the relative boundary of N_F is the disjoint union of the relative interiors of all its faces.

5. Let $c \in \mathrm{cone}\,(a_1, \ldots, a_{m_F})$, so that $c = A_F^\top \lambda$, where $\lambda \in \mathbb{R}^k$ has nonnegative coordinates. Then, for all $x \in F$ and $y \in \mathrm{P} \cap F^c$,

$$\langle c, x \rangle = \langle \lambda, A_F x \rangle = \langle \lambda, b_F \rangle \geq \langle \lambda, A_F y \rangle = \langle A_F^\top \lambda, y \rangle = \langle c, y \rangle$$

since $A_F x = b_F$ and $A_F y < b_F$. Thus, $c \in N_F$ and we have shown that cone $(a_1, \ldots, a_{m_F}) \subseteq N_F$. Conversely, assume that c is a nonzero vector in N_F but $c \notin \mathrm{cone}\,(a_1, \ldots, a_{m_F})$. Then, c is not normal to any supporting hyperplane of F, which implies that there exists a $x \in F$ and $y \in \mathrm{P} \cap F^c$ such that $\langle c, x - y \rangle < 0$, producing a contradiction. Thus, it must be the case that $c \in \mathrm{cone}\,(a_1, \ldots, a_{m_F})$ as well, yielding $N_F \subseteq \mathrm{cone}\,(a_1, \ldots, a_{m_F})$.

$\square$

Appendix C: Checking for the existence of the MLE via Linear Programming.

Deciding whether the MLE exists, that is, whether the vector of observed natural sufficient statistics x is such that $x \in \mathrm{ri}(\mathrm{P})$ is particularly simple if one has access to a $\mathcal{H}$-representation of P as in (1), as indicated in the next result, of immediate verification.

Lemma C1. The MLE exists if and only if the system $Ax \leq b$ is satisfied with strict inequalities.

Unfortunately, this type of representation is typically not available or prohibitively hard to compute, even when k is small, since P may have a number of faces that grow super-exponentially in k (see, for example, Ziegler, 2001).

If instead only a $\mathcal{V}$-representation (3) is available or computable, the existence of the MLE can be established using linear programming, as outlined below. Let B be a matrix whose columns contain the vertices and extreme rays of P, namely the vectors in $\mathcal{Q}$ and $\mathcal{C}$ from equation (3). Then $x \in \mathrm{ri}(\mathrm{P})$ if and only if x can be obtained as a linear combinations of the vectors in $\mathcal{Q}$ and $\mathcal{C}$ with strictly positive coefficients.

Lemma C2. The MLE exists if and only if $x = Bz$, for a vector z with strictly positive coordinates.

This is a feasibility problem which can be decided by solving the linear program

$$
\begin{aligned}
\max\, & s \\
\text{s.t.}\quad Bz\ &=\ x \\
z_i - s\ &\geq\ 0 \\
s\ &\geq\ 0,
\end{aligned}
$$

where z_i denotes the i-th coordinate of z and s is a scalar. If (s^*, z^*) is the optimum, then the MLE exists if and only if $s^* > 0$.

An alternative linear program, which may be computationally preferable, can be formulated based on theorem 4, whose proof can be found in Schrijver (1998), as follows:

$$\max\langle 1, y\rangle$$
$$\text{s.t.} \quad B^\top y = 0$$
$$y \geq 0$$
$$y \leq 1.$$

If y^* is the optimum, the MLE does not exist if and only if $\langle 1, y^*\rangle > 0$.

Theorem C3 (Gordan's Theorem of Alternatives). Given a matrix B, the following are alternatives:

1. $Bx > 0$ has a solution x.
2. $B^\top y = 0$, $y \gneq 0$, has a solution y.

Appendix D: Software

The code used for the analysis and for the figures of the paper is available on the web at

<center>http://www.stat.cmu.edu/~arinaldo/ERG/</center>

The software includes:

1. the `MATLAB` GUI used for creating figures 9, 10 and 11 and some short movies showing the relationship between sequences of natural parameters moving along the outer normals of P and the corresponding sequences of mean values;
2. an MPI `C++` program for complete enumeration of all undirected graphs on n nodes and for counting the number of edges, triangles, k-stars and alternating k-stars. However, complete enumeration is only feasible only for very small graph. Using our program, which can certainly be be improved, it took about 1 hour on a 64-node cluster to enumerate all graphs on 9 nodes, but for the 10-node graph, the estimated running time is about 26.5 days.

5. Acknowledgments

References

BARDORFF-NIELSEN, O. (1978). *Information and Exponential Families in Statistical Theory,* New York: John Wiley & Sons, New York. MR0489333

BARDORFF-NIELSEN, O. (2006). Koopman–Darmois–Pitman Families, *Encycploedia of Statistical Sciences.*

BISHOP, Y.M.M., FIENBERG, S.E., and HOLLAND, P.W. (1975). *Discrete Multivariate Analysis: Theory and Practice.* MIT Press, Cambridge, MA, 1975. Reprinted by Springer-Verlag, New York, 2007. MR2344876

BROWN, L. (1986). *Fundamentals of Statistical Exponential Families*, IMS Lecture Notes-Monograph Series, Vol.9, Hayward, CA.

CSISZÁR, I. and MATÚŠ, F. (2008). Generalized maximum likelihood estimates for exponential families, *Probability Theory and Related Fields*, 141, 213–246. MR2372970

CSISZÁR, I. and MATÚŠ, F. (2005). Closure of exponential families, *Annals of Probability*, 33 (2), 582–600.

CSISZÁR, I. and MATÚŠ, F. (2003). Information projection revisited, *IEEE Transaction of Information Theory*, 49 (6), 1474–1490.

CSISZÁR, I. and MATÚŠ, F. (2001). Convex cores of measures, *Studia Scientiarum Mathematicarum Hungarica*, 38, 177–190. MR1877777

ERIKSSON, N., FIENBERG, S.E., RINALDO, A. and SULLIVANT S. (2006). Polyhedral conditions for the nonexistence of the MLE for hierarchical log-linear models, *Journal of Symbolic Computation*, 41, 222–233. MR2197157

FIENBERG, S.E. and WASSERMAN, S. (1981). An exponential family of probability distributions for directed graphs: comment, *Journal of the American Statistical Association,* 76(373), 54–57. MR0608176

FRANK, O. and STRAUSS, D. (1986). Markov Graphs, *Journal of the American Statistical Association*, 81, 832–842. MR0860518

GEIGER, A., MEEK, C. and STURMFELS, B. (2006). On the toric algebra of graphical models, *Annals of Statistics*, 34(3), 1463–1492. MR2278364

GEYER, C. (2008). Likelihood Inference in Exponential Families and Directions of Recession, Technical Report TR 672, Department of Statistics, University of Minnesota.

GEYER, C.J. and THOMPSON, E.A. (1992). Constrained Monte Carlo maximum likelihood calculations (with discussion), *Journal of the Royal Statistical Society, Series B,* 54, 657–699. MR1185217

GOLDENBERG, A., ZHENG, A.X., FIENBERG, S.E., and AIROLDI, E.M. (2009). A survey of statistical network models, *Foundations and Trends in Machine Learning,* to appear.

GOODREAU, S. (2007). Advances in exponential random graph (p^*) models applied to a large social network, *Social Networks,* 29, 231–248.

HABERMAN, S.J. (1981). An exponential family of probability distributions for directed graphs: comment, *Journal of the American Statistical Association,* 76(373), 60–61. MR0608176

HANDCOCK, M.S., HUNTER, D., BUTTS, C., GOODREAU, S. and MORRIS, M. (2006). Statnet: An R Package for the Statistical Analysis and Simula-

tion of Social Networks. Manual. University of Washington, http://www.csde.washington.edu/statnet.

HANDCOCK, M.S. (2003). Assessing degeneracy in statistical models for social networks. Working paper 39, Center for Statistics and the Social Sciences, University of Washington.

HOLLAND, P.W. and LEINHARDT, S. (1981). An exponential family of probability distributions for directed graphs (with discussion). *Journal of the American Statistical Association* 76 (373), 33–65. MR0608176

HUNTER, D.R., GOODREAU, S.M. and HANDCOCK, M.S. (2008). Goodness of fit of social network models, Journal of the American Statistical Association, 103(481), 248–258. MR2394635

LEE, Y. and NELDER, J.A. (1996). Hierarchical generalized linear models (with discussion), *Journal of the Royal Statistical Society, Series B*, 58, 619–678. MR1410182

LETAC, G. (1992). *Lectures on Natural Exponential Families and Their Variance Functions*, Monografias de Matemática 50. Instituto de Matemática Pura e Aplicada, Rio de Janeiro, Brazil. MR1182991

MORENO, J.L. (1934). *Who Shall Survive?* Nervous and Mental Disease Publishing Company, Washington, DC.

RINALDO, A. (2006a). On maximum likelihood estimation in log-linear models, Technical Report 833, Department of Statistics, Carnegie Mellon University.

RINALDO, A. (2006b). Computing Maximum Likelihood Estimates in Log-Linear Models, Technical Report 835, Department of Statistics, Carnegie Mellon University.

ROBINS, G., PATTISON, P., KALISH, Y. and LUSHER, D. (2007). An introduction to exponential random graph (p^*) models for social networks, *Social Networks,* 29, 171–191.

ROBINS, G, SNIJERS, T., WANG, P., HANDCOCK, M. and PATTISON, P. (2007). Recent developments in exponential random graph (p^*) models for social networks, *Social Networks,* 29, 192–215.

ROCKAFELLAR, R.T. (1970). *Convex Analysis*, Princeton University Press, Princeton. MR0274683

SCHRIJVER, A. (1998). *Theory of Linear and Integer Programming*, Wiley & Sons, New York. MR0874114

SNIJDERS, T.A.B. (2002). Markov Chain Monte Carlo estimation of exponential random graph models, *Journal of Social Structure,* 3.

STRAUSS, D. AND M. IKEDA (1990). Pseudolikelihood estimation for social networks, *Journal of the American Statistical Association,* 85, 204–212. MR1137368

STURMFELS, B. (1995). *Gröbner Bases and Convex Polytopes*, American Mathematical Society, Providence, RI.

WASSERMAN, S. and PATTISON, P.E. (1996). Logit models and logistic regressions for social networks: I. An introduction to Markov graphs and p^*, *Psychometrika* 61, 401–425. MR1424909

WASSERMAN, S.S. and ROBINS, G.L. (2004). An Introduction to Random Graphs, Dependence Graphs, and p^*, in *Models and Methods in Social Net-*

work Analysis, eds. P. Carrington, J.S. and Wasserman, S.S., Cambridge: Cambridge University Press, New York.

ZIEGLER, G.M. (2001). Lectures on 0/1 polytopes, in Polytopes—Combinatorics and Computations, DMS Seminar, Band 29, G. Kalai and G.M. Ziegler editors, Birkhäuser. MR1785291

ZIEGLER, G.M. (1995). *Lectures on Polytopes*, Springer-Verlag, New York. MR1311028