

---

# How Does Maximum Likelihood Estimation for the $p_1$ Model Scale for Large Sparse Networks?

---

Alessandro Rinaldo and Stephen E. Fienberg

Department of Statistics  
Carnegie Mellon University  
Pittsburgh, PA 15213  
arinaldo@cmu.edu, fienberg@stat.cmu.edu

Sonja Petrović

Department of Statistics  
Penn State  
University Park, PA 16802  
sonja@psu.edu

## Abstract

We study the crucial problem of MLE existence for the directed random graph model proposed by Holland and Leinhardt [3], known as the  $p_1$  model, which describes dyadic interactions in a social network. We show that necessary and sufficient conditions for the existence of the MLE are based on polyhedral conditions, which offer a systematic way to characterize (in a combinatorial fashion) sample points leading to a nonexistent MLE, and non-estimability of the probability parameters under a nonexistent MLE. While specifics of nonexistence are best understood for networks of relatively small size, there are important implications of our results for large networks, since current algorithms and software for fitting  $p_1$  and other exponential random graph models (ERGMs) have no simple mechanism for detecting non-existence and identifying non-estimable parameters.

Existence of the maximum likelihood estimator (MLE) is a central issue in the theory and practice of exponential families: statistical inference becomes ill-defined and unfeasible under a nonexistent MLE ([4,5]), as only certain combinations of the true parameters are actually estimable. For ERGMs, studying nonexistence of the MLE has yielded significant insights into the properties of these models such as *degeneracy* [2,4]. For more specialized network models based on *linear* exponential families, this theory can be further refined and has significant algorithmic implication: see, e.g. [5], for the case of the  $\beta$ -model.

In this article we are concerned with the problem of nonexistence of the MLE for the  $p_1$  model of Holland and Leinhardt ([3]), a very popular class of statistical models for analyzing social networks. By representing the  $p_1$  as a linear exponential family, we are able to characterize in a geometric fashion for which patterns of observed networks the MLE of the model parameter exists and which model parameters are estimable under a nonexistent MLE. Unlike the  $\beta$ -model, the  $p_1$  turns out to have much more complicated geometric properties which makes the task of characterizing all instances of network configurations leading to a nonexistent MLE more challenging.

The Holland and Leinhardt  $p_1$  model [3] focuses on the occurrences of directed and bi-directed edges in a random graph representing the interactions among the units of a network. For every pair of nodes  $(i, j)$  we define the probability vector

$$p_{i,j} = (p_{i,j}(0,0), p_{i,j}(1,0), p_{i,j}(0,1), p_{i,j}(1,1)) \quad (1)$$

containing the probabilities of the four possible edge types:  $p_{i,j}(1,0)$ ,  $p_{i,j}(0,1)$  and  $p_{i,j}(1,1)$  denote the probabilities of the edge configurations  $i \rightarrow j$ ,  $i \leftarrow j$  and  $i \leftrightarrow j$ , respectively, and  $p_{i,j}(0,0)$  is the probability that there is no edge between  $i$  and  $j$  in the network. (Thus, 1 denotes the outgoing side of an edge.) By symmetry,  $p_{i,j}(a,b) = p_{j,i}(b,a)$ , for all  $a, b \in \{0,1\}$ , and the four probabilities for each dyad sum to 1. The  $\binom{n}{2}$  dyads are modelled as mutually independent draws from multinomial distributions with class probabilities  $p_{i,j}$ ,  $i < j$ . Specifically, the  $p_1$  model specifies the

multinomial probabilities of each dyad  $(i, j)$  in logarithmic form as follows:

$$\begin{aligned}
\log p_{i,j}(0, 0) &= \lambda_{ij} \\
\log p_{i,j}(1, 0) &= \lambda_{ij} + \alpha_i + \beta_j + \theta \\
\log p_{i,j}(0, 1) &= \lambda_{ij} + \alpha_j + \beta_i + \theta \\
\log p_{i,j}(1, 1) &= \lambda_{ij} + \alpha_i + \beta_j + \alpha_j + \beta_i + 2\theta + \rho_{i,j}.
\end{aligned} \tag{2}$$

The parameter  $\alpha_i$  quantifies the effect of an outgoing edge from node  $i$ , the parameter  $\beta_j$  instead measures the effect of an incoming edge into node  $j$ , while  $\rho_{i,j}$  controls the added effect of reciprocated edges (in both directions). The parameter  $\theta$  measures the propensity of the network to have edges and, therefore, controls the ‘‘density’’ of the graph. The parameters  $\{\lambda_{i,j} : i < j\}$  are normalizing constants to ensure that the 4 probabilities for each each dyad  $(i, j)$  sum to 1, and need not be estimated. Note that, in order for the model to be identifiable, additional linear constraints need to be imposed on its parameters (see [3]). Different variants of the  $p_1$  model can be obtained by constraining the model parameters, e.g., (1)  $\rho_{ij} = 0$ , *no reciprocal effect*; (2)  $\rho_{ij} = \rho$ , *constant reciprocation*; (3)  $\rho_{ij} = \rho + \rho_i + \rho_j$ , *edge-dependent reciprocation*.

Our data take the form of a single observed network. Thus, each dyad  $(i, j)$  is observed in only one of its four possible states and this one observation is a random vector in  $\mathbb{R}^4$  with a Multinomial(1,  $p_{i,j}$ ) distribution. As a result, data are sparse and, even though the dyadic probabilities are strictly positive according to the defining equations (2), only some of the model parameters may be estimated from the data. Extension to the case in which the dyads are observed multiple times are straightforward.

For a network on  $n$  nodes, we represent the vector of  $2n(n-1)$  dyadic probabilities as

$$p = (p_{12}, p_{13}, \dots, p_{n-1,n}) \in \mathbb{R}^{2n(n-1)},$$

where, for each  $i < j$ ,  $p_{ij}$  is given as in (1). The  $p_1$  model is the set of all probability distributions that satisfy the Holland-Leinhardt model (2). By construction, the  $p_1$  model is log-linear or toric: each set of probabilities  $p$  satisfying the  $p_1$  model is such that  $\log p$  is in the linear space spanned by the rows of some *design matrix*  $A$ , which can be constructed as follows (this construction is by no means unique and leads to rank-deficient matrices, though it is rather simple). The columns of  $A$  are indexed by the entries of the vectors  $p_{i,j}$ ,  $i < j$ , where the  $p_{i,j}$ 's are ordered lexicographically, and its rows by the model parameters, ordered arbitrarily. The  $(r, c)$  entry of  $A$  is equal to the coefficient of the  $c$ -th parameter in the logarithmic expansion of the  $r$ -th probability as indicated in (2). In particular, notice that the entries of  $A$  can only be 0, 1 or 2. For example, in the case  $\rho_{ij} = \rho + \rho_i + \rho_j$ , the matrix  $A$  has  $\binom{n}{2} + 3n + 2$  rows. When  $n = 3$ , the design matrix corresponding to this model is

	$p_{1,2}$				$p_{1,3}$				$p_{2,3}$			
$\lambda_{12}$	1	1	1	1	0	0	0	0	0	0	0	0
$\lambda_{13}$	0	0	0	0	1	1	1	1	0	0	0	0
$\lambda_{23}$	0	0	0	0	0	0	0	0	1	1	1	1
$\theta$	0	1	1	2	0	1	1	2	0	1	1	2
$\alpha_1$	0	1	0	1	0	1	0	1	0	0	0	0
$\alpha_2$	0	0	1	1	0	0	0	0	0	1	0	1
$\alpha_3$	0	0	0	0	0	0	1	1	0	0	1	1
$\beta_1$	0	0	1	1	0	0	1	1	0	0	0	0
$\beta_2$	0	1	0	1	0	0	0	0	0	0	1	1
$\beta_3$	0	0	0	0	0	1	0	1	0	1	0	1
$\rho$	0	0	0	1	0	0	0	1	0	0	0	1
$\rho_1$	0	0	0	1	0	0	0	1	0	0	0	0
$\rho_2$	0	0	0	1	0	0	0	0	0	0	0	1
$\rho_3$	0	0	0	0	0	0	0	1	0	0	0	1

Let  $\theta = \{\lambda_{i,j}, \alpha_i, \beta_j, \theta, \rho, \rho_i, \rho_j, i < j\}$  be the vector of model parameters, whose dimension  $d$  depends on the type of model restrictions. Then, using the design matrix  $A$ , one can readily verify that the model has the log-linear representation  $\log p = A^T \zeta$ . Equivalently,

$$p_{i,j}(a, b) = e^{\lambda_{ij}} (e^{\alpha_i})^a (e^{\alpha_j})^b (e^{\beta_i})^b (e^{\beta_j})^a (e^{\theta})^{a+b} (e^{\rho_{ij}})^{\min(a,b)}, \quad \forall i < j, \quad \forall a, b \in \{0, 1\}. \tag{3}$$

We denote with  $M_A \subset \{0, 1\}^{2n(n-1)}$  the set of all such edge probabilities and with  $\mathcal{I}$  the coordinate set of each point in  $M_A$ . Notice that  $\mathcal{I}$  is comprised by  $\binom{n}{2}$  4-tuples indexed by  $((i, j) : i < j)$ , with the elements of the  $(i, j)$ th tuple being the coordinates of the vector (1).

Let  $\mathcal{X}_n \subset \{0, 1\}^{2n(n-1)}$  denote the sample space, i.e. the set of all observable networks on  $n$  nodes. Then, for every  $x \in \mathcal{X}_n$ ,  $x = (x_{1,2}, x_{1,3}, \dots, x_{n-1,n})$ , where each of the  $\binom{n}{2}$  subvectors  $x_{i,j}$  is a vertex of  $\Delta_3$ , the 4-simplex. In fact  $\mathcal{F}$  is also the coordinate set of  $x$ . This way of representing

a bidirected network on  $n$  nodes with a highly-constrained 0/1 vector of dimension  $2n(n-1)$  may appear cumbersome and redundant, especially compared to the original parametrization of [3], requiring only  $\frac{n(n-1)}{2}$  bits. The reason for our choice is simple and important: only with this redundant parametrization the model with the reciprocity parameter becomes a linear exponential family.

For any point  $x \in \mathcal{X}_n$ , the *likelihood function* is the function  $\ell: M_A \rightarrow [0, 1]$  given by

$$\ell_x(p) = \prod_{f \in \mathcal{F}} p(f)^{x(f)}, \quad (4)$$

and the *maximum likelihood estimate* (MLE) of  $p^0$  is  $\hat{p} = \operatorname{argmax}_{p \in M_A} \ell_x(p)$ . Despite  $\ell_x$  being smooth and concave on its domain for each  $x \in \mathcal{X}_n$ , there exist points  $x \in \mathcal{X}_n$  for which the unique supremum of  $\ell_x$  is achieved on the boundary of  $M_A$ , and thus it will have some zero coordinates. In this case, the MLE of  $p_0$  does not exist. Indeed, if a vector  $p$  with zero entries is to satisfy equations (3), then some of the entries in  $\theta$  must be equal to  $-\infty$ . While it is easy to see that zero entries in the vector of sufficient statistics  $Ax$  will lead to such a problem, our analysis below shows that there are many more patterns of zero entries in the vector of edge counts which also cause the MLE to be undefined.

Following the analysis in [5] (see [1] for details), the convex support for this family is the polytope obtained as the Minkowski sum  $P_A := \sum_{i < j} A_{i,j}$ , where  $A_{i,j}$  is the sub-matrix of  $A$  comprised by the four columns referring to the dyad  $(i, j)$ .

**Theorem 0.1.** *The MLE of a  $p_1$  model with design matrix  $A$  exists if and only if  $Ax$  is in the interior of  $P_A$ . A point  $Ax$  belongs to the interior of some face  $F$  of  $P_A$  if and only if there exists a set  $\mathcal{F} \subset \mathcal{I}$  such that*

$$Ax = Ap, \quad (5)$$

where  $p$  is the closure of  $M_A$  and is such that  $p(f) \in \{0, 1\}$  if  $f \notin \mathcal{F}$  and  $p(f) \in (0, 1)$  otherwise. The set  $\mathcal{F}$  is uniquely determined by the face  $F$  and is the maximal set for which (5) holds.

The sets  $\mathcal{F}$  are called *facial sets* and give a combinatorial representation of the boundary (i.e. the faces) of  $P_A$ . The above theorem states that, given an observed network  $x \in \mathcal{X}_n$  the MLE of the parameters exists if and only if  $Ax \in \operatorname{ri}(\mathcal{X}_n)$  and when the MLE does not exist because  $Ax$  belongs to the interior of some face  $F$  of  $P_A$ , only the probability parameters in the corresponding facial set  $\mathcal{F}$  are estimable. Thus, the collections of the complements of all facial sets describe precisely the set of all possible non-estimable network probabilities as across all possible instances of non-existent MLEs. Furthermore, to characterize the network configurations  $x$  for which the MLE fails to exist, one must study the faces of  $P_A$ , a daunting task whose combinatorial complexity is very high and rapidly increasing with  $n$ .

The following lemma shows that one can always replace  $P_A$  with the larger but simpler polyhedron  $C_A := \operatorname{cone}(A)$  to derive its face lattice. This result is of practical importance because the face lattice of  $C_A$  can be handled by algorithms that scale better, as we show in our numerical experiments.

**Lemma 0.1.** *For any  $p_1$  model with associated design matrix  $A$ , the MLE exists if and only if  $Ax \in \operatorname{ri}(C_A)$ , and the facial sets of  $P_A$  are also facial sets of  $C_A$ .*

## Numerical Experiments

We describe some numerical experiments illustrating the reduction in complexity given by Lemma 0.1. Table 1 displays the number of vertices of the polytopes  $P_A$  for the three  $p_1$  model specifications and various network sizes. The last column of the table contains the number of columns of the design matrix, which is also the number of extreme rays of the marginal cone  $C_A$ . In comparison, the number of vertices of  $P_A$ , whose determination is computationally very hard, is very large and grows extremely fast with  $n$ .

In Table 2 we report the number of facets, dimensions and ambient dimensions of the cones  $C_A$  for different values of  $n$  and for the three specifications of the reciprocity parameters  $\rho_{i,j}$ . This provides

$n$	$\rho_{i,j} = 0$	$\rho_{i,j} = \rho$	$\rho = \rho_i + \rho_j$	$2n(n-1)$
3	62	62	62	12
4	1,862	2,415	3,086	24
5	88,232	158,072	347,032	40

Table 1: Number of vertices for the polytopes  $P_A$  for different specifications of the  $p_1$  model and different network sizes. Computations carried out using `minksum`. The last column indicates the number of columns of the design matrix  $A$ , which correspond to the number of generators of  $C_A$ .

$n$	$\rho_{i,j} = 0$			$\rho_{i,j} = \rho$			$\rho = \rho_i + \rho_j$		
	Facets	Dim.	Ambient Dim.	Facets	Dim.	Ambient Dim.	Facets	Dim.	Ambient Dim.
3	30	7	9	56	8	10	15	10	13
4	132	12	14	348	13	15	148	16	19
5	660	18	20	3,032	19	21	1,775	23	26
6	3,181	25	27	94,337	26	28	57,527	31	34

Table 2: No. of facets, dimensions and ambient dimensions of the cones  $C_A$  for different specifications of  $p_1$  and different network sizes. The no. of facets of  $C_A$  equals the no. of facets of  $P_A$  plus  $\binom{n}{2}$ . These additional facets corresponding to the sampling constraints of one observation per dyad.

an indirect measure of the complexity of these models and of the non-zero patterns in extended MLEs, and shows how quickly the complexity of  $p_1$  models may scale with the network size  $n$ .

**The case  $n = 3$ .** The sample space consists of  $4^3 = 64$  possible networks. When  $\rho_{i,j} = 0$  for all  $i$  and  $j$ , there are 63 different observable sufficient statistics, but only one belongs to  $\text{ri}(P_A)$  and thus leads to the existence of the MLE. This sufficient statistic corresponds to the two networks

$$\begin{bmatrix} \times & 0 & 1 \\ 1 & \times & 0 \\ 0 & 1 & \times \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} \times & 1 & 0 \\ 0 & \times & 1 \\ 1 & 0 & \times \end{bmatrix}.$$

where the associated MLE is a 12-dimensional vector whose entries are all 0.25. The polytope  $P_A$  has 62 vertices and 30 facets. When  $\rho_{i,j} = \rho \neq 0$  or  $\rho_{i,j} = \rho_i + \rho_j$  the MLE never exists.

**The case  $n = 4$ .** The sample space contains 4096 observable networks. If  $\rho_{i,j} = 0$ , there are 2,656 different observable sufficient statistics, only 64 of which yield existent MLEs. Overall, out of the 4,096 possible networks, only 426 have MLEs. When  $\rho_{i,j} = \rho \neq 0$ , there are 3,150 different observable sufficient statistics, only 48 of which yield existent MLEs. Overall, out of the 4,096 possible networks, only 96 have MLEs. When  $\rho_{i,j} = \rho_i + \rho_j$ , there are 3,150 different observable sufficient statistics and the MLE never exists.

**The case  $n = 5$ .** The sample space consists of  $4^{10} = 1,048,576$  different networks. If  $\rho_{i,j} = 0$ , there are 225,025 different sufficient statistics, and the MLE exists for 7,983. If  $\rho_{i,j} = \rho \neq 0$  the number of distinct sufficient statistics is 349,500, and the MLE exists in 12,684 cases. Finally, when  $\rho_{i,j} = \rho_i + \rho_j$ , there are 583,346 different sufficient statistics and the MLE never exists.

While specifics of nonexistence are best understood for networks of relatively small size, there are important implications of our results for large networks, since current algorithms and software for fitting  $p_1$  and other exponential random graph models (ERGMs) have no simple mechanism for detecting non-existence and focusing in on those parameters in the model that are estimable. This is separate from the degeneracy problems well-documented for ERGMs, cf. [2,4]. We believe that the nonexistence problem is likely to be especially severe for some forms of ERGMs in common use among social network practitioners.

## References

- [1] Barndorff-Nielsen, O.E. (1978). *Information and Exponential Families in Statistical Theory*, Wiley, New York.
- [2] 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, Seattle.

- [3] Holland, P.W. and Leinhardt, S. (1981). An exponential family of probability distributions for random graphs, *Journal of the American Statistical Association*, 76(373), 33–50.
- [4] Rinaldo, A., S. Fienberg, and Y. Zhou (2009). On the geometry of discrete exponential families with application to exponential random graph models. *Electronic Journal of Statistics* 3, 446–484.
- [5] Rinaldo, A. Petrović S. and Fienberg, S. E. (2012). Maximum likelihood estimation in the  $\beta$ -model, currently under revision for the *Annals of Statistics*.
- [6] Rinaldo, A. Petrović S. and Fienberg, S. E. (2012). Maximum likelihood estimation in network models, available at <http://arxiv.org/abs/1105.6145>