

Conditional Marginalization for Exponential Random Graph Models

Tom A.B. Snijders *

January 21, 2010

To be published, *Journal of Mathematical Sociology*

*University of Oxford and University of Groningen; this paper was written while being a visiting professorial fellow at the University of Melbourne. I thank Pip Pattison and Garry Robins for stimulating discussions; and an anonymous reviewer for comments that led to clarifications.

Conditional Marginalization for Exponential Random Graph Models

Abstract

For exponential random graph models, under quite general conditions, it is proved that induced subgraphs on node sets disconnected from the other nodes still have distributions from an exponential random graph model. This can help in the theoretical interpretation of such models. An application is that for saturated snowball samples from a potentially larger graph which is a realization of an exponential random graph model, it is possible to do the analysis of the observed snowball sample within the framework of exponential random graph models without any knowledge of the larger graph.

Keywords. Connected component, network delineation, network boundary, random graphs, snowball sample.

1 Exponential Random Graph Models

Markov graphs, a class of probability distributions for graphs, were proposed by Frank and Strauss (1986). This was generalized to exponential random graph (p^*) models by Frank (1991) and Wasserman and Pattison (1996). Specifications of these models that made them more widely applicable in practice were proposed by Snijders et al. (2006), and some of the wider applications of these new specifications were presented in Robins et al. (2007).

To define these distributions, consider a finite node set \mathcal{N} and denote the set of all graphs on \mathcal{N} by $\mathcal{Y}(\mathcal{N})$. The term ‘graphs’ here refers to so-called simple graphs, i.e., nondirected graphs without loops and parallel edges; the extension to directed graphs is straightforward. A graph y is a combination of

a node set \mathcal{N} and an edge set $\mathcal{E}(y)$ which is a set of unordered pairs of elements of \mathcal{N} . The edge set will be denoted here by the edge indicator functions y_{ij} , where $y_{ij} = 1$ denotes that there is an edge between nodes i and j – i.e., $\{i, j\} \in \mathcal{E}(y)$ – while $y_{ij} = 0$ denotes that there is no such edge. The matrix $(y_{ij})_{i,j \in \mathcal{N}}$ is the *adjacency matrix* of the graph. The exponential random graph model (ERG model, ERGM) is defined by a probability function of the form

$$P_\theta\{Y = y\} = \exp(\theta'u(y) - \psi(\theta)) \quad y \in \mathcal{Y}(\mathcal{N}) \quad (1)$$

where y is the graph, $u(y)$ is a p -dimensional vector of statistics of the graph, and θ is a p -dimensional parameter. The function $\psi(\theta)$ takes care of the normalization requirement that the probabilities sum to 1. The nodes are supposed to be labeled and there may be covariates defined on the nodes, or pairs of nodes, on which $u(y)$ can also depend.

This paper is concerned with the distributions of graphs on smaller node sets that are induced by exponential random graph models. For a subset \mathcal{N}_1 of \mathcal{N} , we denote the induced subgraph of y on the node set \mathcal{N}_1 by $y|_{\mathcal{N}_1}$. Thus, $y|_{\mathcal{N}_1}$ has node set \mathcal{N}_1 and edge set $\mathcal{E}(y|_{\mathcal{N}_1}) = \{\{i, j\} \in \mathcal{E}(y) \mid i, j \in \mathcal{N}_1\}$. The starting point of this paper is the observation, known since Frank and Straus (1986), that if Y has an ERG distribution (1) and $\mathcal{N}_1 \subset \mathcal{N}$, the induced subgraph $Y|_{\mathcal{N}_1}$ does not in general have an ERG distribution. Thus, if the network delineation (Laumann, Marsden, and Prensky, 1983) would have left out a few nodes, then the remaining observed graph would not have followed an exponential random graph model.

This issue can be seen in the light of the general question for statistical models of what would happen if only part of the data were observed. This is called *marginalization*, because what happens then is determined by the marginal distribution of the observed data. Three examples are the following. For independent identically distributed (i.i.d.) samples from some distribution, if we observe the same variables for a random subsample instead of the whole sample, then still the assumption is valid that we have an i.i.d. sample from this distribution. For a sample from a multivariate normal distribution, if we observe only a subset of the variables then the basic type of assumption

remains valid: multivariate normality for some random vector implies multivariate normality for a subvector. On the other hand, if we consider a model of linear regression with two independent variables X_1 and X_2 and normally distributed i.i.d. residuals where X_2 is a dichotomous variable, then dropping X_2 from the observations destroys the basic properties of the standard linear regression model, as the residuals are not normally distributed any more – e.g., they could have a bimodal distribution. In all these examples, restricting the observation to only part of the data leads to a loss of information; in only the third example the loss of data also hurts by destroying the validity of the basic model assumptions. We can express this by saying that the model *marginalizes* under the loss of observations in the first two cases, but not in the last case. Marginalization implies that if we would observe only part of the variables, the statistical data analysis followed would still be compatible with the analysis of the larger data set in the sense that the model assumptions for the larger data set imply the same type of model assumptions for the reduced data set. Marginalization is regarded as a valuable kind of consistency of a model: if the research design would be such that only part of the data had been observed, still the same kind of statistical analysis would be appropriate.

Exponential random graph models do not marginalize when dropping some nodes from the graph, in the following sense. If Y is a random graph on a node set \mathcal{N} with probability distribution (1), then for a fixed subset $\mathcal{N}_1 \subset \mathcal{N}$, the induced subgraph $Y|_{\mathcal{N}_1}$ will not in general have a probability distribution of this form. Thus, the class of exponential random graph models is not closed under the operation of deleting nodes from the graph. The only known exceptions are trivial models, where edge indicators Y_{ij} are independent. This lack of marginalization has been regarded by some as a defect for the intuitive interpretation of this model: if the specification of the node set would have been different, then the validity of a probability distribution of type (1) would be lost.

This paper treats a kind of marginalization that does hold for exponential random graph models. Section 2 prepares the stage by defining the

requirement of component independence for exponential random graph models, which is a quite natural and broad requirement. Section 3 gives a basic marginalization property for exponential random graph models that holds if this requirement is satisfied. This property can be helpful theoretically for the interpretation of exponential random graph models. A number of corollaries is given to illuminate its consequences. A practical consequence is indicated for the analysis of saturated snowball samples (cf. Doreian and Woodard, 1994), drawn from a larger network distributed according to an exponential random graph model. Such samples can again be analyzed using an exponential random graph model, without requiring information about the rest of the graph and without needing special methods for missing data. In the discussion section it is argued that such a conditional marginalization is indeed more in line with what should be expected for models for network analysis than unconditional marginalization, and the results of the paper are discussed in the context of network delineation.

2 Component Independence

Definition (1) is extremely general as it allows any statistic $u(y)$. In practice, the statistics used for ERGMs are often chosen so as to satisfy certain conditional independence assumptions, such as the Markov dependence assumption (Frank and Strauss, 1986) or the social circuit dependence assumption (this term was coined by Robins et al., 2007, for the assumption used in Snijders et al., 2006). Here we introduce a weak conditional independence assumption which restricts $u(y)$ in a way that will be reasonable in many cases.

Recall that a component, or connected component, of a graph is a maximal connected subgraph. If the graph is such that nodes in \mathcal{N}_1 have no edges to nodes outside \mathcal{N}_1 , i.e.,

$$i \in \mathcal{N}_1, j \in \mathcal{N} \setminus \mathcal{N}_1 \Rightarrow y_{ij} = 0, \quad (2)$$

then the subgraph $y|_{\mathcal{N}_1}$ must be a *union of components* of y .

The new conditional independence is called *component independence*; the interpretation is that dependence occurs only within components, not between components. To define this formally, recall that a *partition* of \mathcal{N} is a set of disjoint subsets $\mathcal{N}_1, \dots, \mathcal{N}_H$, which jointly cover the whole node set:

$$\mathcal{N} = \cup_{h=1}^H \mathcal{N}_h ; \text{ and } \mathcal{N}_h \cap \mathcal{N}_k = \emptyset \text{ for } h \neq k .$$

Component independence is defined as follows.

Definition.

An exponential random graph model on the node set \mathcal{N} is *component independent* if, for every two-subset partition $(\mathcal{N}_1, \mathcal{N}_2)$ of \mathcal{N} , conditional on the event that there are no edges between \mathcal{N}_1 and \mathcal{N}_2 as represented by (2), the induced subgraphs $Y|_{\mathcal{N}_1}$ and $Y|_{\mathcal{N}_2}$ are stochastically independent.

This property is equivalent to the similar requirement for an arbitrary number of components, as stated in the next proposition. This proposition also specifies an equivalent condition in terms of the function $u(y)$ which implicitly specifies the distributions for the induced subgraphs.

Proposition.

Suppose that the ERGM defined by $u(y)$ is component independent, and let $\mathcal{N}_1, \dots, \mathcal{N}_H$ be a partition of \mathcal{N} . Then for any graph y that has no edges between \mathcal{N}_h and \mathcal{N}_k for any $h \neq k$, $u(y)$ can be written as

$$u(y) = \sum_{h=1}^H u(\pi_h(y)) + u_d , \tag{3a}$$

where $\pi_h(y)$ denotes, for each h , the graph on \mathcal{N} that has the same edges as y on \mathcal{N}_h , and no other edges:

$$(\pi_h(y))_{ij} = \begin{cases} y_{ij} & \text{if } i, j \in \mathcal{N}_h \\ 0 & \text{else,} \end{cases} \tag{3b}$$

and where u_d is a constant independent of y .

It is evident that condition (3) also implies component independence, so that this is an equivalent characterization.

Proof.

The proof is by mathematical induction. For a partition $\mathcal{N}_1, \dots, \mathcal{N}_H$, denote by C_H the event that Y that has no edges between \mathcal{N}_h and \mathcal{N}_k for any $h \neq k$. Denote $Y|_{\mathcal{N}_h}$ by Y_h and the empty graph on \mathcal{N}_h by \emptyset_h .

For $H = 2$, the conditional independence of Y_1 and Y_2 implies that

$$\begin{aligned} & \mathbf{P}_\theta\{Y_1 = y_1, Y_2 = y_2 \mid C_2\} \\ &= \frac{\mathbf{P}_\theta\{Y_1 = y_1, Y_2 = \emptyset_2 \mid C_2\} \mathbf{P}_\theta\{Y_1 = \emptyset_1, Y_2 = y_2 \mid C_2\}}{\mathbf{P}_\theta\{Y_1 = \emptyset_1, Y_2 = \emptyset_2 \mid C_2\}} \\ &= k_\theta \mathbf{P}_\theta\{Y = \pi_1(y)\} \mathbf{P}_\theta\{Y = \pi_2(y)\} = \exp(\theta(\pi_1(y) + \pi_2(y)) - k'_\theta) \end{aligned}$$

for constants k_θ, k'_θ independent of y . This implies (3) for $H = 2$. Now suppose that (3) holds for some $H \geq 2$; we shall prove that it holds also for $H + 1$.

Let $\mathcal{N}_1, \dots, \mathcal{N}_{H+1}$ be a partition of \mathcal{N} . Define $\mathcal{N}_+ = \cup_{h=1}^H \mathcal{N}_h$, and define $\pi_+(y)$ as (3b) applied to node set \mathcal{N}_+ . Then $\mathcal{N}_+, \mathcal{N}_{H+1}$ is a partition into two sets, so there is a number u_d^+ such that for any graph y that has no edges between \mathcal{N}_+ and \mathcal{N}_{H+1} , $u(y)$ is equal to

$$u(y) = u(\pi_+(y)) + u(\pi_{H+1}(y)) + u_d^+. \quad (4)$$

Further define $\mathcal{N}_H^* = \mathcal{N}_H = \mathcal{N}_H \cup \mathcal{N}_{H+1}$ and define $\pi_H^*(y)$ as (3b) applied to node set \mathcal{N}_H^* . Consider a graph y satisfying C_{H+1} . By the induction hypothesis applied to the partition $\mathcal{N}_1, \dots, \mathcal{N}_{H-1}, \mathcal{N}_H^*$, we have

$$\begin{aligned} u(\pi_+(y)) &= \sum_{h=1}^{H-1} u(\pi_h(\pi_+(y))) + u(\pi_H^*(\pi_+(y))) + u_d^* \\ &= \sum_{h=1}^H u(\pi_h(y)) + u_d^* \end{aligned} \quad (5)$$

for some u_d^* , where the second equality sign follows from $\pi_h(\pi_+(y)) = \pi_h(y)$ for $h = 1, \dots, H - 1$ and $\pi_H^*(\pi_+(y)) = \pi_H(y)$. Combining (4) and (5) yields

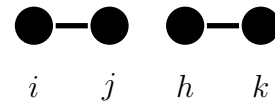
$$u(y) = u(\pi_+(y)) + u(\pi_{H+1}(y)) + u_d^+ = \sum_{h=1}^{H+1} u(\pi_h(y)) + u_d^* + u_d^+.$$

■

Practically all specifications for ERGMs proposed in the literature are component independent. The major example is provided by subgraph counts for connected subgraphs, which are the most widely used statistics because they are the statistics that obey various much stricter conditional independence assumptions, as can be proved from the Hammersley-Clifford theorem (Frank and Strauss, 1986; Pattison and Robins, 2002). For subgraph counts the number u_d is 0. An example where u_d is not zero is the case where $u(y)$ is defined as the number of pairs of nodes that are not reachable from each other. To give an indication of what is excluded by the definition, we give two examples of one-dimensional statistics that do not lead to component independent ERGMs and for which it is easily seen that they do not satisfy a decomposition of the kind (3a).

1. The statistic

$$u(y) = \frac{1}{8} \sum_{i,j,h,k:\{i,j\} \cap \{h,k\} = \emptyset} y_{ij} y_{hk}$$



is a count of subgraphs on four points, composed of two edges involving distinct nodes. The disconnection of the subgraph leads in the ERGM to dependence between tie variables in different components.

2. The statistic

$$u(y) = \sqrt{\sum_{ij} y_{ij}}$$

being a nonlinear function of the edge count, leads to dependence between disconnected parts of the graph.

3 Conditional Marginalization

This section gives a conditional marginalization theorem for component independent ERGMs. The theorem states that under the condition that we observe two or more unions of components, i.e., several mutually disconnected subgraphs, these subgraphs are independent, and again have ERG distributions. Other conditions on the specific subgraphs can be added, e.g., being internally connected, containing a specified number of edges, etc. The theorem can be summarized by saying that for component independent ERGMs, marginalization holds for connected components. The term *conditional* marginalization is used because whether subgraphs are disconnected is itself dependent on the realization of the graph.

In situations where the network represents a system in which interaction or potential influence is indicated by ties, one might say that disconnected subgraphs are subsystems that have nothing to do with each other, and could just as well be studied in mutual isolation. The interpretation of the theorem is that, for component independent ERGMs, these subsystems then indeed can be analysed separately, and using the same ERG model.

Theorem.

Assume that Y has a component independent exponential random graph distribution with sufficient statistic $u(y)$, and let $\mathcal{N}_1, \dots, \mathcal{N}_H$ be a partition of the node set \mathcal{N} . Let A_0 be the event that in Y there are no ties between nodes in \mathcal{N}_h and nodes in \mathcal{N}_k for any $h \neq k$; in other words, that $Y|_{\mathcal{N}_1}, \dots, Y|_{\mathcal{N}_H}$ are unions of components of Y . For $h = 1, \dots, H$, let A_h be events referring only to $Y|_{\mathcal{N}_h}$.

Then conditional on the event $A_0 \cap A_1 \cap \dots \cap A_H$, the subgraphs $Y|_{\mathcal{N}_h}$ for $h = 1, \dots, H$ are independent, and their distributions are given by

$$P_\theta\{(Y|_{\mathcal{N}_h}) = y_h \mid A_h\} = \begin{cases} \exp(\theta' u(\rho_h(y)) - \psi_h(\theta; A_h)) & \text{if } y_h \text{ satisfies } A_h \\ 0 & \text{otherwise,} \end{cases} \quad (6)$$

where $\psi_h(\theta; A_h)$ are normalization constants and where ρ_h is the function

$\rho_h : \mathcal{Y}(\mathcal{N}_h) \rightarrow \mathcal{Y}(\mathcal{N})$ defined for $y \in \mathcal{Y}(\mathcal{N}_h)$ by

$$(\rho_h(y))_{ij} = \begin{cases} y_{ij} & \text{if } i, j \in \mathcal{N}_h \\ 0 & \text{else.} \end{cases} \quad (7)$$

(It may be noted that the functions π_h and ρ_h are similar but formally different: π_h is defined on $\mathcal{Y}(\mathcal{N})$ and deletes all ties outside of \mathcal{N}_h ; ρ_h is defined on $\mathcal{Y}(\mathcal{N}_h)$ and extends a graph on \mathcal{N}_h to a graph with the same edge set but having node set \mathcal{N} .)

Proof.

Denote $A = A_0 \cap A_1 \cap \dots \cap A_H$. Then, for all $y \in \mathcal{Y}(\mathcal{N})$ satisfying condition A ,

$$\begin{aligned} P_\theta\{Y = y \mid A\} &= \frac{\exp(\theta' u(y) - \psi(\theta))}{P\{A\}} \\ &= \exp(\theta' \sum_h u(\pi_h(y)) - c) \end{aligned}$$

for some constant c : the first equality sign holds by definition; the second follows from the Proposition. Now consider a graph $y \in \mathcal{Y}(\mathcal{N})$ satisfying A , and with induced subgraphs $y|_{\mathcal{N}_h} = y_h$. Then $\rho_h(y_h) = \pi_h(y)$ for all h , so that

$$P_\theta\{Y = y \mid A\} = \exp(\theta' \sum_h u(\rho_h(y_h)) - c) .$$

Under condition A there is a one-to-one correspondence between Y and $(Y|_{\mathcal{N}_h}; h = 1, \dots, H)$, so that

$$P_\theta\{Y = y \mid A\} = P_\theta\{Y|_{\mathcal{N}_h} = y_h \text{ for } h = 1, \dots, H \mid A\} .$$

Therefore, for a suitable choice of normalization constants c_h it holds that

$$P_\theta\{Y|_{\mathcal{N}_h} = y_h \text{ for } h = 1, \dots, H \mid A\} = \exp(\theta' \sum_h u(\rho_h(y_h)) - c)$$

which implies that conditional on the event A , the induced subgraphs $Y|_{\mathcal{N}_h}$ are independent and have distributions (7). ■

In the following we give a number of corollaries. For all of them it is assumed that Y has a component independent exponential random graph distribution with sufficient statistic $u(y)$. Note that we here are discussing ERGMs on random node sets, which may be a bit strange at first sight, because normally ERGMs are defined on fixed node sets. But this is not different in principle from what we always have in conditional probability distributions – conditional distributions condition on random events.

In the first two corollaries the events A_h are omitted – or one could say that they are defined as events that are always true. The first corollary is the direct expression of conditional marginalization. If there are two node sets that are not connected by ties, then from the definition we know that the networks on these two node sets are independent; Corollary 1 tells us that each of these networks also has an exponential random graph distribution.

Corollary 1.

If $\mathcal{N}_1 \subset \mathcal{N}$, then conditional on the event that nodes in \mathcal{N}_1 are not linked to nodes outside this set, $Y|_{\mathcal{N}_1}$ has the distribution given by

$$P\{Y|_{\mathcal{N}_1} = y_1\} = \exp(\theta' u(\rho_1(y_1)) - \psi_1(\theta)) , \quad (8)$$

where $\psi_1(\theta)$ is a normalization constant.

Corollary 2 generalizes this to larger numbers of unions of components.

Corollary 2.

If $\mathcal{N}_1, \dots, \mathcal{N}_H$ is a partition of \mathcal{N} , then conditional on the event that in Y there are no ties between nodes in \mathcal{N}_h and nodes in \mathcal{N}_k for $h \neq k$, the induced subgraphs of Y on the node sets \mathcal{N}_h are independent for different h , and their probability distributions are given by

$$P_\theta\{(Y|_{\mathcal{N}_h}) = y_h\} = \exp(\theta' u(\rho_h(y_h)) - \psi_h(\theta)) , \quad (9)$$

where $\psi_h(\theta)$ are normalization constants.

The third corollary can be used for snowball sample designs (Goodman, 1961; Doreian and Woodard, 1994). The saturated snowball sample starting from an initial node set B is the graph induced by the set \mathcal{N}_1 of all nodes i which are either themselves elements of B , or reachable by a path originating in B . Such a path is defined as a sequence of nodes j, i_1, \dots, i_K, i (with $K \geq 0$) where $j \in B$, and all subsequent nodes are linked:

$$Y_{ji_1} = Y_{i_1i_2} = \dots = Y_{i_Ki} = 1.$$

Corollary 3.

If B is a non-empty subset of \mathcal{N} , then conditional on the event that $Y|_{\mathcal{N}_1}$ is the smallest component of Y containing B , i.e.,

$$\mathcal{N}_1 = B \cup \{i \in \mathcal{N} \mid \text{for some } j \in B \text{ there is a path from } j \text{ to } i\}, \quad (10)$$

the induced subgraph $Y|_{\mathcal{N}_1}$ has the distribution given by

$$P\{Y|_{\mathcal{N}_1} = y_1\} = \begin{cases} \exp(\theta'u(\rho_1(y_1)) - c) & \text{if (10) holds} \\ 0 & \text{if (10) does not hold.} \end{cases}$$

This corollary can be used as follows. Suppose we are studying a network for which it is reasonable to assume that it is the outcome of a component independent ERGM. We do not observe the entire graph, but we take a saturated snowball sample starting from an initial node set B , observing all ties adjacent to these nodes and the new nodes to which these ties are also adjacent, and snowballing on until no further nodes are obtained. Then the observed graph, the snowball sample, is the smallest union of components of the network containing all nodes in B . The corollary implies that we can analyze the observed network as an ERGM on the (random) node set observed, as long as we keep into account that it was obtained as a snowball sample. This means that in the Metropolis-Hastings algorithm for generating realizations of the ERGM (cf. Snijders, 2002), the proposal distribution must respect the constraint (10), but nothing else in the spirit of missing data analysis needs to be done.

The fourth corollary implies that if we observe a graph consisting of several connected components, we can analyze those components separately, provided that we respect the condition that they are components. Thus, again, the proposal distribution in the Metropolis-Hastings algorithm has to respect the connectedness of each component.

Corollary 4.

Under the condition that the connected components of Y are defined by the partition $\mathcal{N}_1, \dots, \mathcal{N}_H$, the subgraphs $Y|_{\mathcal{N}_h}$ are independent and have the distributions given by

$$P\{Y|_{\mathcal{N}_h} = y_h\} = \begin{cases} \exp(\theta' u(\rho_h(y_h)) - c_h) & \text{if } y_h \text{ is connected} \\ 0 & \text{if } y_h \text{ is not connected.} \end{cases}$$

Sometimes isolated nodes are left out of a network for further analysis. The fifth corollary shows that this is compatible with an analysis using an ERG model, provided that we take into account the condition that the remaining graph contains no isolated nodes.

Corollary 5.

Under the condition that the isolated nodes of Y are the nodes in the set \mathcal{N}_0 , the subgraph from which the isolates are deleted, $Y|_{\mathcal{N} \setminus \mathcal{N}_0}$, has the distribution given by

$$P\{Y|_{\mathcal{N} \setminus \mathcal{N}_0} = y_1\} = \begin{cases} \exp(\theta' u(\rho(y_1)) - c_1) & \text{if } y_1 \text{ has no isolates} \\ 0 & \text{if } y_1 \text{ has at least one isolated node,} \end{cases}$$

where ρ is defined as (7) for the set $\mathcal{N} \setminus \mathcal{N}_0$.

The sixth corollary links back to what is called the social circuit model in Robins et al. (2007). This is a conditional independence model that requires that given the rest of the graph, edge indicators Y_{ij} and Y_{hk} are independent if the nodes i, j, h, k are all distinct and $Y_{jh}Y_{ik} = Y_{ih}Y_{jk} = 0$; note that the latter condition is equivalent to saying that creating the edges $Y_{ij} = Y_{hk} = 1$ would not lead to a four-cycle through these four nodes. The corollary shows that

component independent ERGMs satisfy a similar implication, obtained by replacing the condition that no four-cycle should be formed by the stronger condition that the two pairs of nodes are not in the same connected component. In other words, component independence is indeed a weaker requirement than the circuit dependence model.

Corollary 6.

For $i, j, h, k \in \mathcal{N}$, denote by $y_{-(ij),-(hk)}$ the adjacency matrix of the graph without the edge indicators y_{ij} and y_{hk} . Assume that $y_{-(ij),-(hk)}$ is such that there is no path from either of the nodes i and j to either of the nodes h and k . In other words, nodes i and j on the one hand, and h and k on the other hand, are in disconnected parts of the graph. Then the random variables Y_{ij} and Y_{hk} are conditionally independent, given $Y_{-(ij),-(hk)} = y_{-(ij),-(hk)}$.

Corollary 6 of course generalizes directly to conditional independence of multiple edge indicators in more than two disconnected subgraphs.

Finally, two corollaries are presented that give conditional distributions of parts of the “small loose objects” remaining outside of the giant component, as often seen in pictures of networks delineated by using a predetermined node set. Here we consider the components of 1, 2, or 3 nodes: there are only four possibilities, viz., isolated nodes, isolated dyads, isolated two-stars, isolated triangles. When we consider the dynamic process that can be employed to construct random draws from ERGMs (Snijders, 2002; Robins et al., 2007), we can see that the total number of such small structures will depend on the parameters that determine how larger structures are formed and connect to smaller structures. However, these corollaries tell us that the *relative* numbers of these four small isolated structures are totally determined by the parameters in the model for small subgraphs: isolates, edges, two-stars, and triangles.

Corollary 7.

Let N_0 be the number of nodes of degree 0 or 1. Suppose that all elements of the sufficient statistic $u(Y)$ are connected subgraph counts, and denote the coefficient of the number of isolates by θ_I and the coefficient of the number of edges, $\frac{1}{2} \sum_{i,j} y_{ij}$, by θ_E . Then, conditional on N_0 , the number of isolated dyads D has probability function

$$\mathbf{P}\{D = d\} = \frac{N_0(N_0 - 1) \dots (N_0 - 2d)}{2^d d!} \exp((\theta_E - 2\theta_I)d - \psi(\theta_I, \theta_E, N_0)) \quad (11)$$

for a normalization constant $\psi(\theta_I, \theta_E, N_0)$.

Proof.

Let \mathcal{N}_0 be the set of nodes of degree 0 or 1. The induced subgraph on \mathcal{N}_0 must consist of D isolated dyads and $N_0 - 2D$ isolates. Within \mathcal{N}_0 , no other connected subgraphs are possible under the assumed condition. Therefore other subgraph counts cannot contribute to this conditional probability, and each induced subgraph on \mathcal{N}_0 has a probability proportional to $e^{(\theta_E - 2\theta_I)d}$. The number of ways of selecting d isolated dyads among N_0 nodes is

$$\frac{N_0(N_0 - 1) \dots (N_0 - 2d)}{2^d d!}.$$

Together, these observations prove (11). ■

Corollary 8.

Let N_0 be the number of isolated 3-node connected subgraphs; note that such subgraphs must be isolated twopaths or isolated triangles. Suppose that the sufficient statistic $u(Y)$ is composed only of connected subgraph counts, and denote the coefficient of the number of edges by θ_E , the coefficient of the number of two-stars by θ_{S_2} , and the coefficient of the number of triangles by θ_T . Then, conditional on N_0 , the number of isolated triangles has a binomial distribution with binomial denominator N_0 and probability parameter

$$\frac{\exp(\theta_E + 2\theta_{S_2} + \theta_T)}{1 + \exp(\theta_E + 2\theta_{S_2} + \theta_T)}. \quad (12)$$

Proof.

We use the theorem, applied to \mathcal{N}_0 being defined as the nodes in isolated 3-node connected subgraphs (which contains $3\mathcal{N}_0$ nodes). The induced subgraph on \mathcal{N}_0 consists of only, and of all, isolated two-stars and isolated triangles. Other subgraph counts cannot play a role for the probability of this induced subgraph. Each isolated two-star contributes $2\theta_E + \theta_{S_2}$ to the exponent. Each isolated triangle contributes $3\theta_E + 3\theta_{S_2} + \theta_T$. The sum of the number of isolated two-stars and isolated triangles is fixed. Hence the relative contribution of isolated triangles with respect to isolated two-stars is $\exp(\theta_E + 2\theta_{S_2} + \theta_T)$. ■

4 Discussion

This paper establishes a conditional marginalization property for a broad class of exponential random graph models (ERGMs), viz., models where mutually disconnected parts of the graph are independent. The latter condition rules out ‘action at a distance’ and is quite natural. The conditional marginalization property states that for such models, the distribution of the graph restricted to a subset of the nodes, under the condition that this subgraph is disconnected from the rest of the graph, still follows an exponential random graph model. This property can be regarded as a support for the theoretical consistency of the ERGM.

To discuss the interpretation of this property let us return to the reasons why in general the validity of marginalization, as it holds, e.g., for the multivariate normal distribution, is a valued property of a statistical model. This property implies that if we would observe only part of the variables, the statistical data analysis followed would of course be less informative because of the loss of data, but compatible with the analysis of the larger data set in the sense that for the reduced data set the same type of model assumptions (in the example: multivariate normality) hold as for the larger data set. For network analysis, however, this is not at all a kind of compatibility that should be expected when nodes are dropped from the network. The delineation of a

network, i.e., the specification of the node set, is an essential first step of network analysis, treated in the literature as the ‘network boundary problem’ (Laumann, Marsden, and Prensky, 1983; Doreian and Woodard, 1994; Marsden 2005). No network analyst would think that arbitrarily deleting nodes from a network would leave the subsequent data analysis still compatible with what it would have been to begin with. Networks are regarded approximately as closed systems (e.g., Doreian and Woodard, op. cit., p. 273) and this basic feature will potentially be violated by deleting nodes from the network. Therefore, it is natural that marginalization of the ERG family of distributions holds for connected components but not for subgraphs induced by arbitrary subsets of nodes.

Several consequences of this marginalization property were presented. Of these consequences, Corollary 3 can have practical importance because it shows that network delineation by a saturated snowball sample design is compatible with analysis by an ERGM. Under the assumption that the snowball sample is carried out in a graph which is the outcome of a component independent ERGM, we do not need any information about the number of nodes outside the snowball sample or the ties between them, and the analysis can be carried out as a regular ERGM analysis of the observed network provided only that in the analysis the extra condition is respected that the observed network was obtained from a snowball sample, as represented in (10).

References

- Doreian, P., and Woodard, K. (1994). Defining and locating cores and boundaries of social networks. *Social Networks*, 16, 267–293.
- Frank, O. (1991). Statistical analysis of change in networks. *Statistica Neerlandica*, 45, 283–293.

- Frank, O., and Strauss, D. (1986). Markov graphs. *Journal of the American Statistical Association*, 81, 832–842.
- Goodman, L.A. 1961. Snowball sampling. *Annals of Mathematical Statistics*, 32, 148–70.
- Laumann, E.O., Marsden, P.V., and Prensky, D. (1983). The boundary specification problem in network analysis. In: Burt, R.S., and Minor, M.J., *Applied Network Analysis*, pp. 18–34. Beverly Hills: Sage.
- Marsden, P.V. (2005). Recent developments in network measurement. In: Carrington, P.J., Scott, J., and Wasserman, S., editors. *Models and methods in social network analysis*, pp. 8–30. Cambridge: Cambridge University Press.
- Pattison, P.E., and Robins, G.L. (2002). Neighbourhood based models for social networks. *Sociological Methodology*, 22, 301–337.
- Robins, G.L., Snijders, T.A.B., Wang, P., Handcock, M., and Pattison, P.E. (2007). Recent developments in exponential random graph (p^*) models for social networks. *Social Networks*, 29, 192–215.
- Snijders, T.A.B. 2002. Markov chain Monte Carlo estimation of exponential random graph models. *Journal of Social Structure*, Vol. 3 (2002), No. 2. <http://www2.heinz.cmu.edu/project/INSNA/joss/index1.html>.
- Snijders, T.A.B., Pattison, P.E., Robins, G.L., and Handcock, M.S. (2006). New specifications for exponential random graph models. *Sociological Methodology*, 26, 99–153.
- Wasserman, S., and Pattison, P.E. 1996. Logit models and logistic regression for social networks: I. An introduction to Markov graphs and p^* . *Psychometrika*, 61, 401–425.