Sampling Networks from Their Posterior Predictive Distribution

RAVI GOYAL, VICTOR DE GRUTTOLA
Department of Biostatistics, Harvard School of Public Health and
JOSEPH BLITZSTEIN
Department of Statistics, Harvard University
(*e-mail:* rgoyal@hsph.harvard.edu)

1 Appendix A: Characterization of Valid Degree Mixing Matrices

Theorem: *Let $TDMM_{i,j}$ represent the number of edges connecting nodes of degree i to nodes of degree j. Let $TD_i = (\sum_j TDMM_{i,j} + TDMM_{i,i})/i$ (this will represent the number of nodes with degree i). A square matrix, $TDMM$, of dimension r is graphical by a simple undirected network if and only if the following five conditions are met.*

1. *TD_i is a non-negative integer*
2. *$TDMM_{i,j} \leq TD_i * TD_j$ if $i \neq j$*
3. *$TDMM_{i,i} \leq TD_i * (TD_i - 1)/2$*
4. *$TDMM_{i,j} \geq 0$*
5. *$TDMM_{i,j} = TDMM_{j,i}$ (symmetric)*

Before we can prove the theorem, we first need the following lemma.

Lemma: *Let $|E| \in \{0, \cdots, n*(n-1)/2\}$ where n is the number of nodes in a graph. The degree sequence d where $d_i \in \{\alpha, \alpha+1\}$ and $d_i \geq 0$ for all $i \in \{1, \cdots, n\}$ and $\sum_{i=0}^{n} d_i = 2*|E|$ is graphical.*

Proof of Lemma: Proof by strong induction on $|E|$.

Base Case: $|E| = 1$. Thus, d has size, n, ≥ 2 and is a set of $n-2$ 0's and exactly two 1's. d is clearly graphical by creating n nodes with the last two having an edge between them.

Induction Step: Assume true for $|E| \leq N$ show for $|E| = N+1$. Let d be a degree sequence of size n, where $d_i \in \{\alpha, \alpha+1\}$ and $d_i \geq 0$ for all $i \in \{1, \cdots, n\}$, $\sum_{i=0}^{n-1} d_i = 2*(N+1)$ and $N+1 \in \{0, \cdots, n*(n-1)/2\}$. Let $M = min\{i : d_i \geq d_j$ for all $j \in \{1, \cdots, n\}\}$. Now, construct a degree sequence d'; initially let d' be equal to d. Next subtract one from the largest d_M values in d', excluding position M; therefore, $d'_{i_j} = d_{i_j} - 1$ for all $j \in \{1, \cdots, d_M\}$ where $\{d'_{i_1}, \cdots, d'_{i_k}\}$ are the largest d_M values in d' (excluding position M). Lastly, remove node M from d'; therefore, d' is of size $n-1$. This construction is possible because $d_M = \lceil \frac{2*(N+1)}{n} \rceil \leq \frac{n*(n-1)}{n} = n-1$ and $d_{i_j} > 0$ for all $j \in \{1, \cdots, d_M\}$ since if $d_M = 1$ then $\exists$ a $j \neq M$ such that $d_j = 1$ because $\sum_{i=0}^{n-1} d_i$ is even.

In order to check if d' is graphical, we need to ensure $\frac{\sum_{i=0}^{n-1} d'_i}{2} \in \{0, \cdots, (n-1)*(n-2)/2\}$ and $d'_i \in \{\alpha, \alpha+1\}$ for all $i \in \{1, \cdots, n-1\}$. By assumption we know that $N+1 \leq \frac{n*(n-1)}{2}$. Thus, it can be shown that $N+1-\frac{2*(N+1)}{n} \leq \frac{(n-1)*(n-2)}{2}$. Since $N+1-\lceil\frac{2*(N+1)}{n}\rceil = \frac{\sum_{i=0}^{n-1} d'_i}{2}$ we get the desired result that $\frac{\sum_{i=0}^{n-1} d'_i}{2} \in \{0, \cdots, (n-1)*(n-2)/2\}$. $d'_i \in \{\alpha', \alpha'+1\}$ for all $i \in \{1, \cdots, n-1\}$ is guaranteed since we are subtracting one for the degrees with the highest values and d originally had the property that $d_i \in \{\alpha, \alpha+1\}$ for all $i \in \{1, \cdots, n\}$.

With these two conditions met, we can use the induction assumption, and thus d' is graphical. Including an isolate node at position M would still make the sequence graphical. Finally, connecting the isolate node to $\{d'_{i_1}, \cdots, d'_{i_k}\}$ would still be graphical. This new graph would have the degree sequence of d, and so d is graphical.

□

Proof of Theorem

Given an undirected graph, it is clear that the degree mixing matrix will satisfy the conditions in Theorem 1.1. Thus, we need only show that a matrix which satisfies the five criteria is graphical, which will be shown by constructing a realization of the matrix. We begin by generating an empty network with $\sum_i TD_i$ nodes, where TD_i of them will have degree i. The first condition guarantees that TD_i is a non-negative integer. The next step is adding edges to the empty graph. This will be separated into two steps. The first step is adding edges between nodes with the same final degree and the second step is adding edges between nodes with different final degrees.

Step 1: Edges between nodes with same final degree

The goal of step one is to connect $TDMM_{i,i}$ edges between nodes with final degree i for each $i \in \{1, \cdots, r\}$. We want to connect the edges such that at the end of this step each node of final degree i has one of two possible degree values, $\lfloor\frac{TDMM_{i,i}}{TD_i}\rfloor$ and $\lceil\frac{TDMM_{i,i}}{TD_i}\rceil$, for its current degree, i.e. the edges are added to balance the current degree as much as possible. The assignment of TD_i ensures that maximum degree after this step, $\lceil\frac{TDMM_{i,i}}{TD_i}\rceil$, is less than or equal to the desired final degree, i. In order to prove that edges can be added to maintain the required degree balance we use the lemma, where $TDMM_{i,i}$ and TD_i represent $|E|$ and n respectively. To apply the lemma, we need to insure $TDMM_{i,i} \in \{0, \cdots, \frac{(TD_i*(TD_{i-1}))}{2}\}$, which is guaranteed by conditions (3) and (4).

Step 2: Connect nodes with different degrees

Once edges have been added to nodes with the same final degree, we have to add edges between nodes of degree i to nodes of degree j, for each $i, j \in \{1, \cdots, r\}$. Define the following for each i, j pair where $i \neq j$. Let $\vec{\alpha}_i$ denote a vector where the k^{th} term, α_{i_k}, equals i minus current degree of the k^{th} node with degree i, i.e. the number of edges still needed for each node. Similarly, define $\vec{\alpha}_j$ for nodes with degree j. Without loss of generality we assume that $\vec{\alpha}_i$ and $\vec{\alpha}_j$ are in decreasing order. Define $\vec{\beta}_i$ such that $\beta_{i_k} \in$

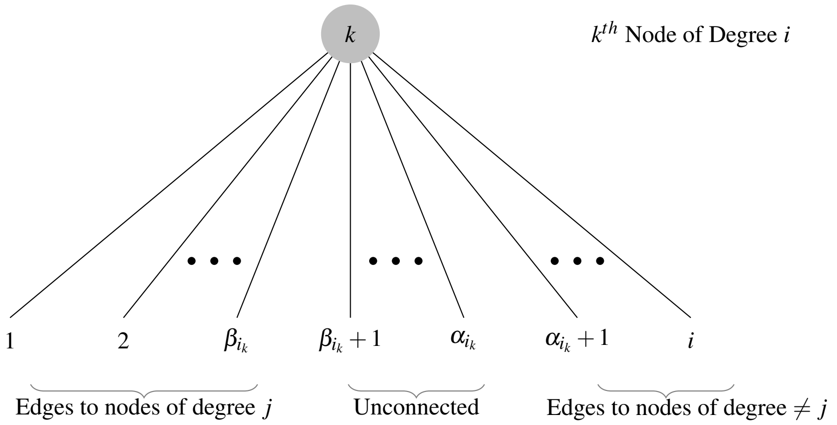
k
k^{th} Node of Degree i
1
2
β_{i_k}
$\beta_{i_k}+1$
α_{i_k}
$\alpha_{i_k}+1$
i
Edges to nodes of degree j
Unconnected
Edges to nodes of degree $\neq j$

Fig. 1. Edge connections for a node of degree i

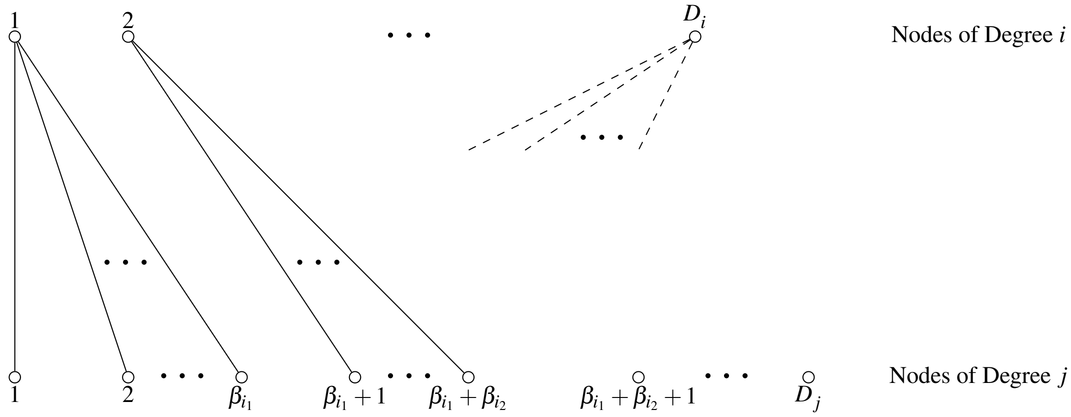

Fig. 2. Edge connections between nodes of degree i and degree j.

$\{\lfloor\frac{TDMM_{i,j}}{TD_i}\rfloor,\lceil\frac{TDMM_{i,j}}{TD_i}\rceil\}$, $\sum_k \beta_{i_k} = TDMM_{i,j}$, and $\beta_{i_1} \geq \beta_{i_2} \geq \cdots \geq \beta_{i_{TD_i}}$. β_{i_k} represents the number of edges that will be added which connect the k^{th} node with degree i with nodes of degree j. Similarly, define $\vec{\beta}_j$ for nodes with degree j. Figure 1 graphical describes the edge connections for a node of degree i.

Connect the first degree i node to the first β_{i_1} nodes of degree j. Next connect the second degree i node to the next β_{i_2} nodes of degree j (may need to loop back to the first degree j node). This process is described in figure 2.
Repeat this process for all TD_i degree i nodes. This process can fail in one of three ways to construct a graph with the degree mixing matrix of $TDMM$.

Issue 1: $\beta_{i_k} > TD_j$.

The issue 1 occurs when a single node, k, of degree i must connect to β_{i_k} nodes of degree j, but β_{i_k} is greater than the number of nodes of degree j, TD_j. Thus, node k must form two edges with the same node of degree j. This cannot occur because $\beta_{i_k} \leq \lceil \frac{TDMM_{i,j}}{TD_i} \rceil \leq TD_j$ by conditions (1) and (2).

Issue 2: $\alpha_{i_k} < \beta_{i_k}$.

The second issue occurs when $\alpha_{i_k} < \beta_{i_k}$, i.e. a node of degree i has fewer unconnected edges than the number of nodes of degree j to which it is assigned. Initially when constructing the graph we generated $TD_i = (\sum_j TDMM_{i,j} + TDMM_{i,i})/i$ nodes of degree i, which means the sum degree of all the degree i nodes is $(\sum_j TDMM_{i,j} + TDMM_{i,i})$. The number of unconnected edges after step 1 is $(\sum_j TDMM_{i,j} + TDMM_{i,i}) - 2 * TDMM_{i,i} = \sum_{j \neq i} TDMM_{i,j}$. Thus, there are enough unconnected edges from nodes of degree i to connect the required number of edges, $TDMM_{i,j}$. Hence, we know $\sum_k \alpha_{i_k} - \beta_{i_k} \geq 0$. Thus, there exists partitions, p_1 and p_2, of size TD_i of the values $\sum_k \alpha_{i_k}$ and $\sum_k \beta_{i_k}$ such that $p_{1_l} \geq p_{2_l}$ for each $l \in \{1, \cdots, TD_i\}$. One such pair of partitions is where each partition is decreasing and is as balanced as possible. This is exactly the partition generated under this construction proof. Throughout the construction the number of available edges for nodes with the same degree are as balanced as possible. The first step of connecting edges between nodes with the same degree initially forces this condition. In subsequent steps of connecting nodes with different degrees ensures this condition remains by assigning more edges to those nodes with more available edges. Thus, by construction $\alpha_{i_k} < \beta_{i_k}$ is not possible.

Issue 3: $\alpha_{j_k} < \beta_{j_k}$.

Due to the symmetry of i and j, the proof that $\alpha_{j_k} < \beta_{j_k}$ is not possible is identical to issue 2.

□

2 Appendix B: Additional Simulations

Four simulated datasets each containing a sample of 100 nodal degrees were drawn from a negative binomial, Poisson lognormal, Waring, and Yule distribution with parameter sets $\{v = (5, 0.2), maxdeg = 8\}, \{v = c(0.6, 1.2), maxdeg = 8, cutoff = 0\}, \{v = c(3.5, 0.1), maxdeg = 8\}$, and $\{rho = 4, maxdeg = 8\}$, respectively; see commands in Handcock (2003) for additional details. Table 1 shows the values of Y from the 100 simulated nodal degrees for each of the four simulated datasets.

Using the procedure outlined in section 5.2 to specify the PPND, 50,010,000 networks (10,000 removed for burn-in) were generated for each dataset and every 1,000th network is used for analysis; the generated networks were of size 1000. The marginal plots in figures 3–6 are calculated by computing the degree distribution for each of the 50,000 simulated networks. As in section 5.3, the black and red lines represent the target and simulated PPSNSD.

Table 1. **Sampled Nodal Degrees for Simulated Datasets**

Degree	0	1	2	3	4	5	6	7	8
Negative Binomial	0	24	21	17	10	9	6	9	4
Poisson lognormal	27	23	12	15	12	3	1	3	4
Waring	0	69	14	6	3	4	2	1	1
Yule	0	65	23	5	2	2	0	2	1

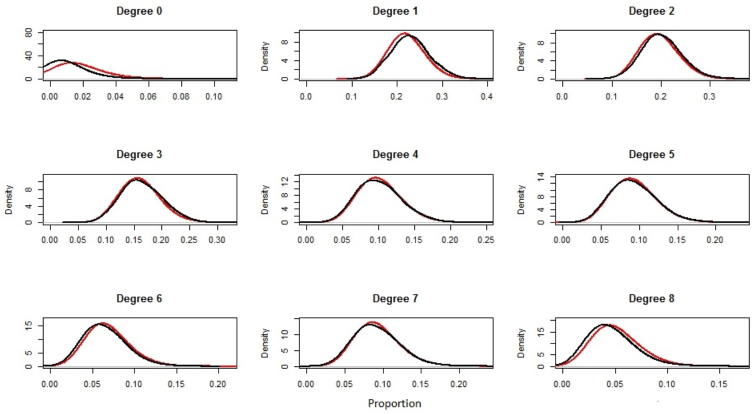

Fig. 3. Negative Binomial: The black lines represent the target PPSNSD. The solid red lines represent the simulated PPSNSD using the proposed methods.

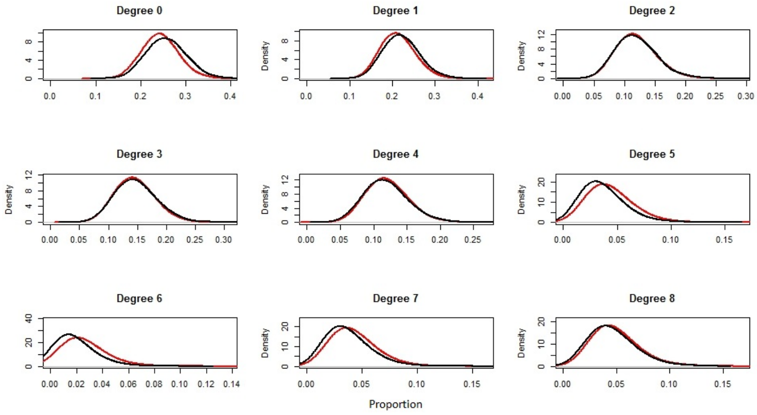

Fig. 4. Poisson Lognormal: The black lines represent the target PPSNSD. The solid red lines represent the simulated PPSNSD using the proposed methods.

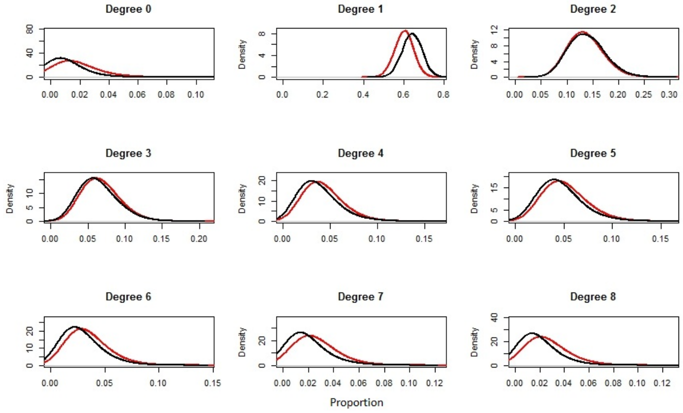

Fig. 5. Waring: The black lines represent the target PPSNSD. The solid red lines represent the simulated PPSNSD using the proposed methods.

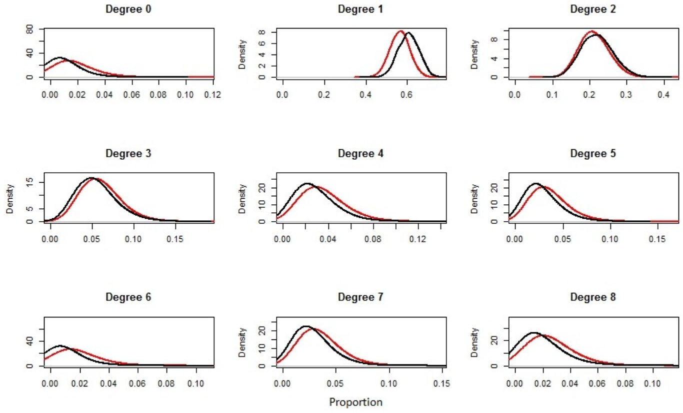

Fig. 6. Yule: The black lines represent the target PPSNSD. The solid red lines represent the simulated PPSNSD using the proposed methods.

Bibliography

Handcock, Mark S. (2003). *degreenet: Models for skewed count distributions relevant to networks*. Seattle, WA. Version 1.2 . Project home page at http://statnet.org.