Statistics and Its Interface Volume 11 (2018) 649–656

Sampling strategies for conditional inference on multigraphs

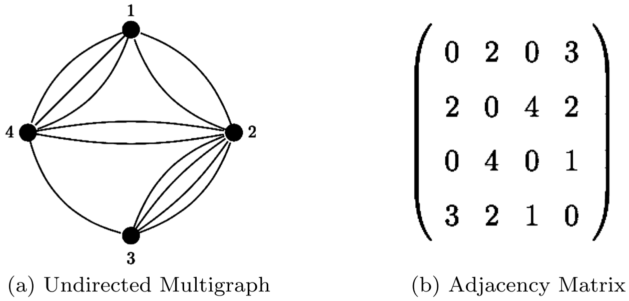
Robert D. Eisinger and Yuguo Chen*,†

We propose two new methods for sampling undirected, loopless multigraphs with fixed degree. The first is a sequential importance sampling method, with the proposal based on an asymptotic approximation to the total number of multigraphs with fixed degree. The multigraphs and their associated importance weights can be used to approximate the null distribution of test statistics and additionally estimate the total number of multigraphs. The second is a Markov chain Monte Carlo method that samples multigraphs based on similar moves used to sample contingency tables with fixed margins. We apply both methods to a number of examples and demonstrate excellent performance.

Keywords and phrases: Counting problem, Exact test, Monte Carlo method, Multigraph, Sequential importance sampling, Symmetric contingency table.

1. INTRODUCTION

Network data is extremely common and there is currently a huge interest in statistical methods for analyzing networks. Fields as diverse as ecology, sociology, and economics deal with networks on a regular basis and require statistical approaches and analysis strategies. Substantial literature is available on methods for graphs with only a single edge between nodes (simple graphs); however, relatively less time has been spent on the case where the network may have multiple links between edges. A network of this type is commonly called a multigraph. Performing statistical inference on multigraphs is of interest to researchers. For example, they may be interested in the number of emails sent between pairs of people in a social group, or the number of interactions observed between pairs of animals. As a small, toy example, consider Figure 1, which shows an undirected, loopless multigraph and the equivalent adjacency matrix. Note that multigraphs may be expressed as graphs with integer-weighted edges, and Figure 1(b) may also be called a weighted adjacency matrix.



$$\begin{pmatrix} 0 & 2 & 0 & 3 \\ 2 & 0 & 4 & 2 \\ 0 & 4 & 0 & 1 \\ 3 & 2 & 1 & 0 \end{pmatrix}$$

(a) Undirected Multigraph

(b) Adjacency Matrix

Figure 1. An undirected multigraph and its associated adjacency matrix.

Researchers are often interested in investigating whether or not a pattern or property of interest on a graph is surprising. For example, one may wish to determine if the number of triplets, the average path length, or some measure of connectedness is unusual. If some pattern or property deviates from random, it may suggest that further study on that property is needed to understand the unusual behavior, and in some cases, these unusual patterns can help us build networks with the desired properties [32]. To detect deviations from randomness in network properties, a test must be performed comparing the observed graph to other possible graphs. The pattern or property of the observed graph is ascertained and compared to the same pattern or property for the comparison graphs. To perform this task, we condition on the degree sequence and consider the observed graph to be a uniform draw from the set of all possible graphs with the same degree sequence. This is an application of exact inference, which requires no potentially inaccurate asymptotic approximations [1, 8, 17]. Conditioning on the degree sequence also creates a probabilistic basis for a test in situations where the subjects were not obtained by sampling but are the only ones available [17].

Several Markov chain Monte Carlo (MCMC) algorithms for sampling simple graphs from the uniform distribution have been proposed [26, 22, 19, 12], and importance sampling methods were considered in [27], [6], [4] and [31]. Reference number [23] developed an effective and widely used method for sampling simple graphs with given degree sequence. Graphs with weighted adjacency matrices have been explored in [30], [29], and [13]. Relatively less attention has been paid to the problem of sampling multigraphs.

To sample multigraphs, we may equivalently sample adjacency matrices. Substantial work has been done on the task of sampling two-way contingency tables with fixed margins. Reference number [7] considered an importance sam-

*Corresponding author.

†Partially supported by the NSF grant DMS-1406455.

pling method, reference number [14] developed a rejection method, and a random walk based MCMC method was discussed in [10]; see [10] for a review. In the case of sampling undirected, loopless multigraphs, the matrices are square, symmetric, and have a zero diagonal.

Here, we are concerned with sampling multigraphs with no self-loops uniformly from the set of all such multigraphs with fixed degree sequence. Based on these sampled graphs, the distribution of a test statistic may be approximated. Additionally, we are interested in estimating the total number of multigraphs with the same fixed degree sequence.

Sampling from the uniform distribution over multigraphs with fixed degree is difficult. Here, we propose a new sequential importance sampling (SIS) method that uses the asymptotic approximation of [5] to guide the sampling. A multigraph is generated and its associated importance weight is used to correct for the bias incurred by sampling. Using these graphs and weights, the distribution of any test statistic may be estimated, and we may additionally obtain an approximation to the number of multigraphs. We also propose an MCMC method for sampling multigraphs with fixed degree.

This paper is organized in the following way: Section 2 introduces the basics of SIS. Section 3 describes how the approximation is incorporated into the proposal to perform SIS. Section 4 proposes an MCMC method for sampling multigraphs. Section 5 provides applications, including an analysis of the clustering of a primate social network and the resilience of an airline network, as well as counting the number of graphs. Section 6 provides concluding remarks.

2. SEQUENTIAL IMPORTANCE SAMPLING

Multigraphs can be expressed equivalently as a symmetric integer-valued adjacency matrix with a zero diagonal, so to sample multigraphs we may equivalently sample adjacency matrices. Let $\Sigma_{\mathbf{d}}$ denote the set of all $n \times n$ symmetric matrices with row margins $\mathbf{d} = (d_1, \ldots, d_n)$, non-negative integer entries, and a zero diagonal, $M = \sum_{i=1}^n d_i$, and $|\Sigma_{\mathbf{d}}|$ the total number of matrices in the set. Denote by T a matrix in $\Sigma_{\mathbf{d}}$, and by $p(T) = 1/|\Sigma_{\mathbf{d}}|$ the uniform distribution over $\Sigma_{\mathbf{d}}$.

If we are interested in estimating $\mu = E_p[f(T)]$, and a matrix $T \in \Sigma_{\mathbf{d}}$ can be simulated from a proposal distribution $q(\cdot)$ that can be easily sampled from and includes the support of $\Sigma_{\mathbf{d}}$, then we may estimate μ using the weighted average of $T_1, \ldots, T_N$, independent and identically distributed (iid) samples drawn from $q(T)$,

$$\hat{\mu} = \frac{\sum_{i=1}^N f(T_i)\frac{p(T_i)}{q(T_i)}}{\sum_{i=1}^N \frac{p(T_i)}{q(T_i)}} = \frac{\sum_{i=1}^N f(T_i)\frac{\mathbb{1}_{\{T_i \in \Sigma_{\mathbf{d}}\}}/|\Sigma_{\mathbf{d}}|}{q(T_i)}}{\sum_{i=1}^N \frac{\mathbb{1}_{\{T_i \in \Sigma_{\mathbf{d}}\}}/|\Sigma_{\mathbf{d}}|}{q(T_i)}} = \frac{\sum_{i=1}^N f(T_i)\frac{\mathbb{1}_{\{T_i \in \Sigma_{\mathbf{d}}\}}}{q(T_i)}}{\sum_{i=1}^N \frac{\mathbb{1}_{\{T_i \in \Sigma_{\mathbf{d}}\}}}{q(T_i)}}. \tag{1}$$

Additionally, the total number of graphs can be written as

$$|\Sigma_{\mathbf{d}}| = \sum_{T \in \Sigma_{\mathbf{d}}} \frac{1}{q(T)} q(T) = E_q\left[\frac{\mathbb{1}_{\{T \in \Sigma_{\mathbf{d}}\}}}{q(T)}\right], \tag{2}$$

so if we are interested in estimating $|\Sigma_{\mathbf{d}}|$, we may use the estimator

$$\widehat{|\Sigma_{\mathbf{d}}|} = \frac{1}{N} \sum_{i=1}^N \frac{\mathbb{1}_{\{T_i \in \Sigma_{\mathbf{d}}\}}}{q(T_i)}. \tag{3}$$

The efficiency of the above estimators may be quantified in several ways. The standard error of $\hat{\mu}$ can be estimated by either repeatedly running the procedure or using an approximation based on the Δ-method [7]:

$$\mathrm{se}(\hat{\mu}) \approx \sqrt{\frac{\mathrm{var}_q\left(\frac{f(T)p(T)}{q(T)} - \mu\frac{p(T)}{q(T)}\right)}{N}}. \tag{4}$$

The *effective sample size*, ESS $= N/(1 + \mathrm{cv}^2)$, is another way to assess method efficiency [16]. Here, the *coefficient of variation* (cv) is given by

$$\mathrm{cv}^2 = \frac{\mathrm{var}_q(p(T)/q(T))}{E_q^2(p(T)/q(T))}. \tag{5}$$

The cv^2 is the χ^2 distance between the proposal and the target, so a small value indicates that the proposal is sampling from a distribution that is close to the desired target. The ESS approximates how many iid samples are equivalent to the N weighted samples obtained through SIS. The theoretical value of cv^2 is unknown, so the sample version is used.

The choice of the proposal $q(\cdot)$ determines the efficiency of the importance sampling procedure. This is a high-dimensional problem, so the strategy that will be employed here is to decompose the proposal into lower dimensional components. The first component of the matrix is sampled, and then the second component of the matrix is sampled conditional on the realization of the first component. The remainder of the matrix is sampled sequentially in a similar way conditional on the realization of all previous components.

3. SAMPLING MULTIGRAPHS

We are proposing a new SIS technique which samples the matrix column by column and uses an asymptotic approximation of [5] to guide the sampling.

If we denote the columns of T by $t_1, \ldots, t_n$, then the probability of sampling a matrix T using a proposal $q(\cdot)$ can be written as

$$q(T = (t_1, \ldots, t_n)) = q(t_1) \times q(t_2|t_1) \ldots q(t_n|t_{n-1}, \ldots, t_1). \tag{6}$$

We begin by sampling the first column of the matrix, t_1, conditional on $\mathbf{d}$. After t_1 has been sampled, the degree sequence is updated, the first column is removed, and we sample the first column of the remaining $(n-1) \times (n-1)$ submatrix. Denote the configuration of the first column by $t_1 = (0, \alpha_{21}, \ldots, \alpha_{n1})$, and denote by $\mathbf{d}^{(2)}$ the updated margins of the $(n-1) \times (n-1)$ submatrix after the first column has been sampled, i.e.,

$$\mathbf{d}^{(2)} = (d_2 - \alpha_{21}, d_3 - \alpha_{31}, \ldots, d_n - \alpha_{n1}). \tag{7}$$

This procedure is repeated until all of the columns have been sampled and a completed matrix is obtained.

We start by writing the true marginal distribution of t_1 under the uniform distribution over $\Sigma_{\mathbf{d}}$. For a given configuration of the first column, $t_1 = (0, \alpha_{21}, \ldots, \alpha_{n1})$, the true marginal distribution of t_1 is

$$p(t_1 = (0, \alpha_{21}, \ldots, \alpha_{n1})) = \frac{|\Sigma_{\mathbf{d}^{(2)}}|}{|\Sigma_{\mathbf{d}}|}. \tag{8}$$

This expression cannot be calculated directly, but an asymptotic formula for $|\Sigma_{\mathbf{d}}|$ was given by [5].

The asymptotic approximation that will be employed was obtained by specializing Theorem 1 of [5] to our setting.

Approximation 1: Given $\mathbf{d} = (d_1, \ldots, d_n)$ and $M = \sum_{i=1}^n d_i$,

$$|\Sigma_{\mathbf{d}}| \sim \Delta_{\mathbf{d}} \equiv \frac{f(M)}{\prod_{i=1}^n d_i!} \exp\{\boldsymbol{a}(\mathbf{d})\}, \tag{9}$$

where $f(M) = M!/[(M/2)!2^{M/2}]$ and $\boldsymbol{a}(\mathbf{d}) = (\sum_i \binom{d_i}{2}/M)^2 - \sum_i \binom{d_i}{2}/M$.

The justification is given in the Appendix. This approximation assumes that all marginal sums are bounded above by a constant d^* and that $M \to \infty$.

The proposal used to sample the first column t_1 is based on Approximation 1 and is shown below. Denote this method SIS-BC.

Proposal 1: The proposal for the first column based on the [5] approximation is

$$\begin{aligned} q(t_1 = (0, \alpha_{21}, \ldots, \alpha_{n1})) \propto \Delta_{\mathbf{d}^{(2)}} \propto \Delta'_{\mathbf{d}^{(2)}} \equiv \\ \frac{1}{\prod_{i=2}^n (d_i - \alpha_{i1})!} \exp\{\boldsymbol{a}(\mathbf{d}^{(2)})\}, \end{aligned} \tag{10}$$

where $\boldsymbol{a}(\cdot)$ is defined as in Approximation 1.

The justification is provided in the Appendix. Although $q(t_1)$ in the above proposal may be sampled directly using enumeration, this is not feasible for larger matrices. In these cases, enumeration takes a long time and it is more convenient to sample $q(t_1)$ using the following rejection method. This is the strategy that will be employed in this paper.

1. Generate a configuration of the first column $\mathbf{a} = (a_1, \ldots, a_n)$ from $g(\mathbf{a})$, where $g(\mathbf{a})$ is the uniform distribution over all possible configurations of the first column. This can be done using the procedure described by [14], which generates length n columns that sum to d_1 until a column is obtained that satisfies the row constraints.
2. Generate a $u \sim$ Unif[0,1].
3. Calculate the ratio $q(\mathbf{a})/(cg(\mathbf{a}))$, where $q(\mathbf{a})$ is the proposal of SIS and c is a constant chosen so that $q(\mathbf{a}) \leq cg(\mathbf{a})$ for any $\mathbf{a}$.
4. Accept $\mathbf{a}$ if $u \leq q(\mathbf{a})/(cg(\mathbf{a}))$. Otherwise, reject $\mathbf{a}$.

Note that $\Delta_{\mathbf{d}^{(2)}}$ will be obtained for every possible configuration of the first column when the normalizing constant for $q(t_1)$ is calculated, so both the number of configurations of the first column and the maximum value of $\Delta_{\mathbf{d}^{(2)}}$ over these configurations are relatively easy to calculate. These quantities may be used to obtain a value c such that $q(\mathbf{a}) \leq cg(\mathbf{a})$ for all $\mathbf{a}$.

3.1 Refined sampling

While the above procedure will yield reasonable estimates, there will be a certain percentage of matrices generated that are invalid. This may occur after some of the columns have been sampled because there is no multigraph that corresponds to the updated degree sequence of the submatrix. Consider the following small example. If the margins are $\{2, 2, 2\}$ and the first column sampled is $t_1 = \{0, 2, 0\}$, then the updated margins for the 2×2 submatrix are $\{0, 2\}$, and the sampling cannot proceed because this degree sequence does not correspond to a valid multigraph. A sequential importance sampling procedure that guarantees the existence of every matrix takes into account an existence condition of [11], cited in [21], to guarantee that every generated matrix is valid. This is the procedure that will be used in Section 5.

Theorem 1. *[11] A degree sequence $d_n \geq d_{n-1} \geq \cdots \geq d_1$ is multigraphical if and only if $\sum_{i=1}^n d_i$ is even and $d_n \leq \sum_{i=1}^{n-1} d_i$.*

This condition is incorporated through an additional rejection step. Only those columns that guarantee the existence of a multigraph are sampled so that the sampling method generates 100% valid matrices. This approach takes longer to run than sampling without generating valid matrices; however, it provides an advantage in terms of cv^2 and standard error, as well as guaranteeing that every generated matrix will be valid.

The SIS procedure for sampling $n \times n$ matrices N times with degree sequence $\mathbf{d}$ and calculating their associated importance weights is described below. We use $\sum_{\mathbf{d}^{(i)} \to \mathbf{d}^{(i+1)}}$ to denote the summation over all possible ways to move from $\mathbf{d}^{(i)} = (d_i^{(i)}, \ldots, d_n^{(i)})$ to $\mathbf{d}^{(i+1)}$ after filling in column t_i. Note that $\mathbf{d}^{(1)} = \mathbf{d}$ and the sampling of t_i is done using the rejection method described in Section 3.

Algorithm 1: SIS-BC for sampling matrices

```
for $j$ in $1, \dots, N$ do
    $w_j = 1$
    for $i$ in $1, \dots, n$ do
        Calculate $S(\mathbf{d}^{(i)}) = \sum_{\mathbf{d}^{(i)} \to \mathbf{d}^{(i+1)}} \Delta'_{\mathbf{d}^{(i+1)}}$
        Sample $t_i$ with probability $\Delta'_{\mathbf{d}^{(i+1)}} / S(\mathbf{d}^{(i)})$ from the set of all possible configurations for the $i^{\text{th}}$ column
        $w_j = w_j \times S(\mathbf{d}^{(i)}) / \Delta'_{\mathbf{d}^{(i+1)}}$
        $\mathbf{d}^{(i+1)} = (d^{(i)}_{i+1} - t_{i,i+1}, \dots, d^{(i)}_{n} - t_{i,n})$
    Output sampled matrix $T_j = (t_1, \dots, t_n)$ with weight $w_j$
```

4. MCMC METHOD

Sampling and testing multigraphs may also be performed using an MCMC procedure based on the [10] method for sampling contingency matrices. At each step, two rows i_1 and i_2 are chosen from $\{1, \dots, n\}$, and two columns j_1 and j_2 are chosen from $\{1, \dots, n\}$, where $j_1 \neq i_1, i_2$ and $j_2 \neq i_1, i_2$. One of the following two moves is made with equal probability on the four cells at the intersection of rows i_1 and i_2 and columns j_1 and j_2:

$$\begin{matrix} +1 & -1 \\ -1 & +1 \end{matrix} \qquad\qquad \begin{matrix} -1 & +1 \\ +1 & -1 \end{matrix}.$$

The same move is then made on the cells opposite the ones sampled to maintain the symmetry constraint. More specifically, the move is performed on both cells (i_1, j_1), (i_2, j_1), (i_2, j_2), (i_1, j_2), and cells (j_1, i_1), (j_1, i_2), (j_2, i_2), (j_2, i_2). If a negative entry is obtained, the new matrix is rejected and the Markov chain stays at the current matrix. As the degree sequence of the matrix is not altered by these moves, every matrix generated will be valid.

Theorem 2. *Choosing two rows i_1 and i_2 from $\{1, \dots, n\}$, and two columns j_1 and j_2 from $\{1, \dots, n\}$, where $j_1 \neq i_1, i_2$ and $j_2 \neq i_1, i_2$, and performing $\begin{smallmatrix} -1 & +1 \\ +1 & -1 \end{smallmatrix}$ or $\begin{smallmatrix} +1 & -1 \\ -1 & +1 \end{smallmatrix}$ with equal probability and the corresponding move on the cells opposite the diagonal constitutes an irreducible Markov Chain on $\Sigma_{\mathbf{d}}$.*

The proof is given in the Appendix and follows [10]. This method has the advantage of being extremely easy to implement. It also allows for the sampling of larger and denser matrices compared to SIS-BC. Although the chain is sticky, in cases where other methods fail it provides a relatively easy way to sample multigraphs.

5. APPLICATIONS AND SIMULATIONS

We illustrate the efficacy of the methods by describing a number of applications and simulations. For SIS-BC, the refined sampling procedure is used in all cases. Computation was performed on a MacBook Pro with a 2.2 GHz Intel Core i7 processor. Coding was done in C with calculation of statistics performed in R.

5.1 Estimating the number of multigraphs

Calculating the total number of multigraphs with a prescribed, fixed degree sequence is difficult. Although only of minor interest as a combinatorial problem, estimating the number of multigraphs provides a good way to test the efficacy of the method. If the SIS procedure estimates the number of multigraphs well in situations where the truth is known, that will provide support for more practical settings where the truth is not known. An exhaustive search is feasible for very small matrices, but will take a prohibitively long time for matrices that are even moderately large. A method to calculate the true number of multigraphs with fixed degree sequence is provided in the free software LattE [3]. The exact number for the three 8×8 matrices provided in Table 1 were calculated using this method in approximately 330 seconds for each multigraph. This method provides the exact answer for small multigraphs, but for larger cases the computation time becomes prohibitive. Using an SIS strategy, we may estimate $|\Sigma_{\mathbf{d}}|$ using (3), based on iid samples from our asymptotically-guided proposal distribution.

We estimate the number of matrices in a few examples. We consider three 8×8 matrices with increasing margins, a 9×9 matrix with all margins equal to 4, a 14×14 matrix with all margins equal to 2, a 26×26 matrix with all margins equal to 5, a 30×30 matrix with all margins equal to 3, and a real 15×15 matrix of chimpanzee grooming behavior [28]. To further test the method we also consider a 20×20 matrix with moderately rough margins equal to $\{15, 5, 5, 5, 5, 5, 5, 5, 5, 5, 1, \dots, 1\}$, a large and sparse 100×100 matrix with all margins equal to 2, and an extremely large and sparse 200×200 matrix with margins equal to $\{3, 1, \dots, 1\}$. The simulation results, along with the exact number of multigraphs calculated by [20] when they are available or using LattE, are given in Table 1. Estimates are based on 1,000 samples and the number following the ± sign denotes the standard error calculated using the Δ-method (4). Simulation results indicate that SIS-BC is performing well in the task of estimating the total number of multigraphs, producing accurate estimates in a relatively short amount of time. We may also conclude that the method is working well even for matrices that are both large and sparse.

Table 1. Performance of SIS-BC for estimating the number of matrices

Matrix	Truth	Estimated # matrices	cv^2	Time (sec)
8×8 with margins $= 2$	$6,202$	$(6.1819 \pm 0.0457) \times 10^3$	0.0546	0.03
8×8 with margins $= 5$	$45,163,496$	$(4.5277 \pm 0.0537) \times 10^7$	0.1404	0.1
8×8 with margins $= 8$	$20,547,642,185$	$(2.0751 \pm 0.0356) \times 10^{10}$	0.2942	0.55
9×9 with margins $= 4$	$170,816,680$	$(1.7468 \pm 0.0199) \times 10^8$	0.1297	0.2
14×14 with margins $= 2$	$10,157,945,044$	$(1.0205 \pm 0.0058) \times 10^{10}$	0.0247	0.2
15×15 chimpanzee data	-	$(1.0089 \pm 0.0543) \times 10^{25}$	2.8990	3.3
20×20 with rough margins	-	$(1.0813 \pm 0.0079) \times 10^{20}$	0.0538	56.5
26×26 with margins $= 5$	1.2836×10^{56}	$(1.2839 \pm 0.0108) \times 10^{56}$	0.0703	386.7
30×30 with margins $= 3$	1.5998×10^{45}	$(1.5890 \pm 0.0080) \times 10^{45}$	0.0253	26.1
50×50 with margins $= 3$	-	$(7.4774 \pm 0.0355) \times 10^{91}$	0.0225	350.1
100×100 with margins $= 2$	-	$(4.1248 \pm 0.0571) \times 10^{56}$	0.0191	≈1hr
200×200 with margins $= \{3, 1, \ldots, 1\}$	-	$(2.1984 \pm 0.0059) \times 10^{188}$	0.0007	≈1hr

Sampling using SIS-BC without guaranteeing validity for the 9×9 matrix with all margins equal to 4 yields an estimate of $(1.7173 \pm 0.0378) \times 10^8$ with cv$^2 = 0.4853$ and 73.2% valid samples in 0.2 seconds. For the 30×30 matrix with all margins equal to 3, an estimate of $(1.5994 \pm 0.0197) \times 10^{45}$ is obtained with cv$^2 = 0.1521$ and 89.5% valid samples in 23.9 seconds.

5.2 Primate social network data

We will consider chimpanzee grooming data collected by [28] in the Budongo Forest in Uganda. This data, pictured in Figure 2, represents a symmetrized version of grooming interactions among fifteen chimpanzees. Each node represents a chimpanzee and the node labels correspond to names. Each link represents a grooming interaction between a pair of chimpanzees, so the counts are the total number of grooming interactions between a given chimpanzee pair. Although the original data was directed, we summed across the diagonal to obtain a symmetric matrix of sociopositive body interactions between pairs of chimpanzees [15]. The diagonal is zero since here chimpanzees are not able to groom themselves.

We are interested in the property of group cohesiveness and specifically would like to investigate whether or not this graph is surprising when considered to be a random draw from the set of all possible networks with fixed degree sequence. This property was considered in the context of primate data in [18]. This will be quantified using the average of the weighted clustering coefficients for each node as defined by [2],

$$C_w = \frac{1}{N} \sum_i c_i^w, \tag{11}$$

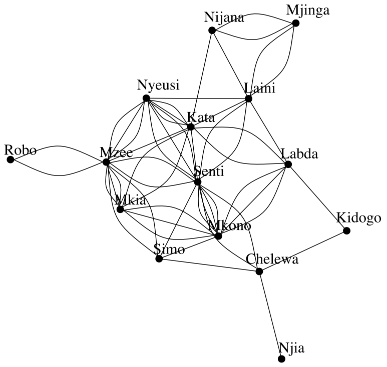


Figure 2. The fifteen node multigraph of chimpanzee grooming relations [28].

where the weighted clustering coefficient for node i is

$$c_i^w = \frac{1}{s_i(k_i - 1)} \sum_{j,h} \frac{(w_{ij} + w_{ih})}{2} a_{ij} a_{ih} a_{jh}. \tag{12}$$

Here $a_{ij} = 1$ if there is an edge between nodes i and j and zero otherwise, w_{ij} is the number of edges between nodes i and j, $s_i = \sum_{j=1}^N a_{ij} w_{ij}$ and $k_i = \sum_j a_{ij}$.

High values of C_w indicate a large degree of overall clustering in the network. The null hypothesis is that C_w is not unusual when the observed network is considered to be a uniform draw from the set of networks with the same degree. We obtain a p-value of 0.5737 ± 0.02797 with cv$^2 = 2.8990$ in a few seconds, indicating no statistical significance. Using the alternative MCMC method for these data with 1,000,000 iterations and 100,000 burn-in, with standard error calculated using the batch means method results in a p-value of 0.5562 ± 0.00108. Although this standard error is smaller than the one obtained using SIS-BC, it appears to be an

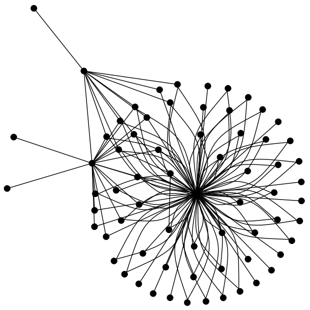


Figure 3. The PSA Airlines network. Nodes represent airports and each edge represents a flight.

underestimate. Running the MCMC procedure 100 times with a different SIS-BC generated starting position each time yields a standard error of 0.03820. It appears that the chain is sticky and takes a long time to explore the space, resulting in an underestimate of the standard error.

5.3 Airline network resilience

The airline network is an important aspect of the national transportation system, responsible for moving millions of passengers every year according to the [24]. History has shown that the airline network is vulnerable to disruption by both targeted attacks and random events. For example, the terrorist attacks in 2001, the eruption of Eyjafjallajökull in 2010, and the United technical glitches in July 2015 all resulted in delays and grounded flights. These consequences impose huge costs on both passengers and the airline industry.

We are interested in investigating the resilience of an airline network to a targeted attack. Here, resilience refers to the ability of the network to maintain short-weighted paths (defined later) between nodes in response to the removal of an important hub airport. For simplicity, we examine only the flights of PSA airlines, a regional airline headquartered in Ohio. Nodes in this network represent airports, and the number of edges between nodes represents the number of flights between the two airports for the month of December 2010 [9]. There are 68 total airports and 135 edges.

To measure the resilience of the network we use the average of the closeness centrality values for all of the nodes [25]. The closeness centrality for a node i is the sum of the inverse of the shortest weighted paths between all other nodes and node i, i.e.,

$$C(i) = \sum_{j:j\neq i} \frac{1}{d(i,j)}, \tag{13}$$

where $d(i,j)$ represents the shortest weighted path between nodes i and j. Because two airports may be considered to be 'close' if they have a large number of flights between them, we define the distance as

$$d(i,j) = \frac{1}{w_{ii_2}} + \frac{1}{w_{i_2i_3}} + \cdots + \frac{1}{w_{i_kj}}, \tag{14}$$

where $\{i, i_2, i_3, \ldots, i_k, j\}$ are the nodes on the shortest weighted path between nodes i and j, and $w_{i_{k-1}i_k}$ is the number of edges between nodes i_{k-1} and i_k.

The overall closeness of the network is the average of the closeness values for all nodes (i.e., $\sum_{i=1}^n C(i)/n$). We are interested in determining if the airline network is less resilient to targeted attacks than would be expected by chance. Eliminating the airport with the second largest degree causes the closeness value to decrease by 2.93%, indicating that it is harder to traverse the network following the removal of the airport with the second largest degree. To test the significance of this change, we generated 1,000 random graphs with the same degree sequence using SIS-BC, eliminated the node corresponding to the airport with the second largest degree and calculated the percent change in closeness. Based on these samples, the probability of seeing a 2.93% or more decrease in average closeness is 0.04427 ± 0.01889, indicating that the airline network formed in such a way that it is less resilient to attacks than we would expect by chance. Computation was completed in under an hour and a cv^2 of 9.4476 was obtained.

6. DISCUSSION

We have developed an SIS strategy for sampling multigraphs with fixed degree based on an asymptotic approximation of [5]. This method samples column by column and performs best in cases where the graph is at least moderately sparse. As the graph becomes denser, performance decreases as judged by cv^2. As the graph becomes larger, the computation time increases, which limits the application of SIS in cases where the graph of interest has a large dimension. Additionally, as the margins increase for a fixed dimension graph, the performance, as judged by cv^2, suffers moderately. Columns may be sampled in any order, but the best performance as judged by cv^2 and standard error and based on simulation is generally given by sampling columns in decreasing order. We have also proposed an MCMC method based on the moves described by [10] for sampling contingency tables. This method performs well even in cases where the graph is extremely dense.

Both methods we have proposed are extremely flexible, as the distribution of any test statistic of interest related to multigraphs may be approximated and a p-value estimated. The SIS method is most suitable for situations where the graph is at least moderately sparse. As the matrix becomes large or dense, the performance and computation time of SIS-BC both suffer, and eventually the use of SIS-BC will become infeasible due to computation time. In these situations, the MCMC method should be preferred, although the standard error as calculated by batch means appears to be an underestimate.

The approximation of [5] may be used to approximate the number of graphs where a set of entries in the adjacency matrix are forced to be structural zeros. This additional feature of the approximation may be leveraged to allow for cell by cell sampling of the adjacency matrix of a multigraph with fixed degree. A cell is first sampled and then forced to be a structural zero. While this approach is appealing, in practice it tends to perform poorly and column by column sampling is preferred. A potential avenue for future work is sampling matrices where any two nodes are connected by no more than r edges.

APPENDIX A

A.1 Justification of Approximation 1

Denote by $(m_{ij})_{n\times n}$ an $n\times n$ symmetric 0-1 matrix, where $m_{ij}=0$ denotes a structural zero at position (i,j). Let $\Delta_{\mathbf{d}}$ be the number of $n\times n$ symmetric matrices over $[0,t]$ such that $g_{ij}=0$ whenever $m_{ij}=0$ and $\sum_j g_{ij}=d_i$. According to Theorem 1 of [5],

$$\Sigma_{\mathbf{d}} \sim \Delta_{\mathbf{d}} \equiv T(M,\delta)\exp\{\epsilon a-b\}/\prod_{j=1}^{n} d_j!,$$

where $M=\sum\limits_{j=1}^{n} d_j$, $\delta=\sum\limits_{m_{ii}=0} d_i$, $\epsilon=1$ if $t>1$ and $\epsilon=-1$ if $t=1$, $a=\left(\sum\limits_{j=1}^{n}\binom{d_j}{2}/M\right)^2$, $b = \left(\sum\limits_{i<j,m_{ij}=0} d_id_j+\sum\limits_{j=1}^{n}\binom{d_j}{2}\right)/M$, $T(M,\delta) = \sum\limits_{j}\binom{M-\delta}{j}C_{M-j}$, and $C_j = j!/((j/2)!2^{j/2})$ if j is even and 0 if j is odd.

In the case of multigraphs with no self-loops, the diagonal is zero and we have $\epsilon=1$, $a=(\sum_{j=1}^{n}\binom{d_j}{2}/M)^2$, $b=\sum_{j=1}^{n}\binom{d_j}{2}/M$, $\delta=M$, and also $T(M,\delta)=M!/((M/2)!2^{M/2})$. Plugging these values in yields the expression in Approximation 1.

A.2 Justification of Proposal 1

The approximation of [5] implies that the number of multigraphs after sampling the first column is approximately

$$(15)\qquad |\Sigma_{\mathbf{d}^{(2)}}| \sim \Delta_{\mathbf{d}^{(2)}} \equiv \frac{f(M-2d_1)}{\prod_{i=2}^{n}(d_i-\alpha_{i1})!}\exp\{\boldsymbol{a}(\mathbf{d}^{(2)})\},$$

and the approximation to the total number of multigraphs $|\Sigma_{\mathbf{d}}|$ is given in Approximation 1. Combining the two expressions above yields the proposal SIS-BC:

$$q(t_1=(0,\alpha_{21},\dots,\alpha_{n2})) \propto \Delta_{\mathbf{d}^{(2)}} \propto \frac{1}{\prod_{i=2}^{n}(d_i-\alpha_{i1})!}\exp\{\boldsymbol{a}(\mathbf{d}^{(2)})\},$$

where $\boldsymbol{a}(\cdot)$ is as in Approximation 1.

A.3 Proof of Theorem 2

We need to show that for every $A,B\in\Sigma_{\mathbf{d}}$, there is a sequence of moves of type

$$\begin{matrix} +1 & -1 \\ -1 & +1 \end{matrix} \qquad\qquad \begin{matrix} -1 & +1 \\ +1 & -1 \end{matrix}$$

leading from A to B. We will use induction.

First define $d(A,B)=\sum_{i,j}|a_{ij}-b_{ij}|$ and note that $d(A,B)$ is divisible by 4 and has minimum nonzero value equal to 8.

Assume the induction hypothesis, that if $0\le d(A,B)\le 4k$ there is a path joining A and B.

This is true for $d(A,B)=8$ because then there are only 8 elements (4 on either side of the diagonal) for which $|a_{ij}-b_{ij}|=1$. Call these cells (i_1,j_1), (i_2,j_1),(i_1,j_2),(i_2,j_2) and (j_1,i_1), (j_2,i_1), (j_1,i_2), (j_2,i_2), where $i_1<i_2$ and $j_1<j_2$. Then we can make an appropriate move to decrease $d(A,B)$ by 8 and obtain $A=B$.

Next, suppose $d(A,B)=4(k+1)$. We will show there is a move from $A\to A'$ where $d(A',B)\le 4k$ or a move $B\to B'$ where $d(A,B')\le 4k$.

Suppose A and B have different elements in the first column (otherwise we can remove the first column and first row and check the second column). Suppose a_{i_11} is the first element at which A and B differ and that $a_{i_11}<b_{i_11}$. Then there exists an i_2 such that $a_{i_21}>b_{i_21}$ and a j_2 such that $a_{i_1j_2}>b_{i_1j_2}$, where $i_2>i_1$ since a_{i_11} is the first element at which A and B differ and $j_2\ne i_1$ since $a_{i_1i_1}$ is a structural zero.

There are two cases. The first is that $i_2\ne j_2$ and the second is that $i_2=j_2$. In both of these cases we will show that there is a move $A\to A'$ where $d(A',B)\le 4k$ or a move $B\to B'$ where $d(A,B')\le 4k$.

In the first case where $i_2\ne j_2$ make the move

$$\begin{matrix} a'_{i_11}=a_{i_11}+1 & a'_{i_1j_2}=a_{i_1j_2}-1 \\ a'_{i_21}=a_{i_21}-1 & a'_{i_2j_2}=a_{i_2j_2}+1 \end{matrix}$$

and the corresponding symmetric move

$$\begin{matrix} a'_{1i_1}=a_{1i_1}+1 & a'_{j_2i_1}=a_{j_2i_1}-1 \\ a'_{1i_2}=a_{1i_2}-1 & a'_{j_2i_2}=a_{j_2i_2}+1 \end{matrix}.$$

We know $a_{i_21},a_{1i_2},a_{i_1j_2},a_{j_2,i_1}>0$ since $a_{i_21}>b_{i_21}$ and $a_{i_1j_2}>b_{i_1j_2}$. Since $i_2\ne j_2$ there are no structural zeros. Moving from $A\to A'$ results in a decrease in the difference of A' with respect to B of 6 on $(i_1,1)$, (i_1,j_2), $(i_2,1)$, $(1,i_1)$, (j_2,i_1), $(1,i_2)$. The difference on (i_2,j_2) and (j_2,i_2) may increase by 2, but the net change is at least 4.

So $d(A',B)\le d(A,B)-4\le 4(k+1)-4=4k$.

In the second case, $i_2=j_2$. Here (i_2,i_2) is a structural zero, so we cannot make any move with rows i_1 and i_2 and columns 1 and i_2.

However, there exists j_2' such that $a_{i_2j_2'}<b_{i_2j_2'}$ where $j_2'\ne 1$ and $j_2'\ne i_1$ because $a_{i_21}>b_{i_21}$ and $a_{i_2i_1}>b_{i_2i_1}$.

Make the below move on B

$$\begin{aligned} b'_{i_1 1} &= b_{i_1 1} - 1 \quad & b'_{i_1 j'_2} &= b_{i_1 j'_2} + 1 \\ b'_{i_2 1} &= b_{i_2 1} + 1 \quad & b'_{i_2 j'_2} &= b_{i_2 j'_2} - 1 \end{aligned}$$

and the corresponding symmetric move

$$\begin{aligned} b'_{1 i_1} &= b_{1 i_1} - 1 \quad & b'_{j'_2 i_1} &= b_{j'_2 i_1} + 1 \\ b'_{1 i_2} &= b_{1 i_2} + 1 \quad & b'_{j'_2 i_2} &= b_{j'_2 i_2} - 1 \end{aligned} .$$

Moving from $B \to B'$ results in $d(A, B') \leq d(A, B) - 4 \leq 4k$. The case where $a_{i_1 1} > b_{i_1 1}$ is symmetric, simply reverse roles of A and B.

Received 15 January 2017

REFERENCES

[1] Agresti, A. (1992). *Categorical Data Analysis*. Wiley, New York. MR1044993

[2] Barrat, A., Barthélemy, M., Pastor-Satorras, R. and Vespignani, A. (2004). The architecture of complex weighted networks. *Proceedings of the National Academy of Sciences* **101** 3747–3752.

[3] Barvinok, A. I. (1994). A Polynomial Time Algorithm for Counting Integral Points in Polyhedra When the Dimension is Fixed. *Mathematics of Operations Research* **19** 769–779.

[4] Bayati, M., Kim, J. H. and Saberi, A. (2010). A sequential algorithm for generating random graphs. *Algorithmica* **58** 860–910. MR2726458

[5] Bender, E. A. and Canfield, E. R. (1978). The asymptotic number of labeled graphs with given degree sequences. *Journal of Combinatorial Theory, Series A* **25** 296–307. MR0505796

[6] Blitzstein, J. and Diaconis, P. (2010). A sequential importance sampling algorithm for generating random graphs with prescribed degrees. *Internet Mathematics* **6** 489–522.

[7] Chen, Y., Diaconis, P., Holmes, S. P. and Liu, J. S. (2005). Sequential Monte Carlo methods for statistical analysis of tables. *Journal of the American Statistical Association* **100** 109–120.

[8] Cochran, W. G. (1952). The χ^2 test of goodness of fit. *The Annals of Mathematical Statistics* **23** 315–345.

[9] Csardi, G. (2014). igraphdata: a collection of network data sets for the igraph package R package version 0.2.

[10] Diaconis, P. and Gangolli, A. (1995). Rectangular arrays with fixed margins. In *Discrete Probability and Algorithms* (D. Aldous, P. Diaconis, J. Spencer and J. Steele, eds.) Springer-Verlag, New York.

[11] Hakimi, S. L. (1962). On realizability of a set of integers as degrees of optimally edge-connected multigraphs. *Journal of the Society for Industrial and Applied Mathematics* **10** 496–506.

[12] Handcock, M. S., Hunter, D. R., Butts, C. T., Goodreau, S. M. and Morris, M. (2008). statnet: software tools for the representation, visualization, analysis and simulation of network data. *Journal of Statistical Software* **24** 1–11.

[13] Hillar, C. and Wibisono, A. (2013). Maximum entropy distributions on graphs. *arXiv preprint* **arXiv:1301.3321**.

[14] Holmes, R. B. and Jones, L. K. (1996). On uniform generation of two-way tables with fixed margins and the conditional volume test of Diaconis and Efron. *The Annals of Statistics* **24** 64–68.

[15] Kasper, C. and Voelkl, B. (2009). A social network analysis of primate groups. *Primates* **50** 343–356.

[16] Kong, A., Liu, J. S. and Wong, W. H. (1994). Sequential imputations and Bayesian missing data problems. *Journal of the American Statistical Association* **89** 278–288.

[17] Lehmann, E. L. (1959). *Testing Statistical Hypotheses*. Wiley, New York. MR0107933

[18] Lehmann, L. and Boesch, C. (2009). Sociality of the dispersing sex: the nature of social bonds in West African female chimpanzees, *Pan* troglodytes. *Animal Behavior* **77** 377–387.

[19] McDonald, J. W., Smith, P. W. F. and Forster, J. J. (2007). Markov Chain Monte Carlo exact inference for social networks. *Social Networks* **29** 127–136.

[20] McKay, B. D. and McLeod, J. C. (2012). Asymptotic enumeration of symmetric integer matrices with uniform row sums. *Journal of the Australian Mathematical Society* **92** 367–384.

[21] Meierling, D. and Volkmann, L. (2008). A remark on the degree sequences of multigraphs. *Mathematical Methods of Operations Research* **69** 369–374.

[22] Milo, R., Shen-Orr, S., Itzkovitz, S., Kashtan, N., Chklovskii, D. and Alon, U. (2002). Network motifs: simple building blocks of complex networks. *Science* **298** 824–827.

[23] Molloy, M. and Reed, B. (1996). A critical point for random graphs with a given degree sequence. *Random Structures & Algorithms* **6** 161–180.

[24] Bureau of Transportation Statistics (2015). Airline Activity: National Summary (U.S. Flights).

[25] Opsahl, T., Agneessens, F. and Skvoretz, J. (2010). Node centrality in weighed networks: Generalizing degree and shortest paths. *Social Networks* **32** 245–251.

[26] Roberts, J. M. (2000). Simple methods for simulation sociomatrices with given marginal totals. *Social Networks* **22** 273–283.

[27] Snijders, T. A. B. (2006). Enumeration and simulation methods for 0-1 matrices with given marginals. *Psychometrika* **56** 397–417.

[28] Sugiyama, Y. (1969). Social behavior of chimpanzees in the Budongo forest, Uganda. *Primates* **10** 197–225.

[29] Yan, T., Qin, H. and Wang, H. (2016). Asymptotics in undirected random graph models parameterized by the strengths of vertices. *Statistica Sinica* **26** 273–293.

[30] Yan, T., Zhao, Y. and Qin, H. (2015). Asymptotic normality in the maximum entropy models on graphs with an increasing number of parameters. *Journal of Multivariate Analysis* **133** 61–76. MR3282018

[31] Zhang, J. and Chen, Y. (2013). Sampling for conditional inference on network data. *Journal of the American Statistical Association* **108** 1295–1307.

[32] Zhang, J. and Chen, Y. (2015). Exponential random graph models for networks resilient to targeted attacks. *Statistics and Its Interface* **8** 267–276.

Robert D. Eisinger
Department of Mathematics, Statistics
and Computer Science
St. Olaf College
Northfield, MN 55057
USA
E-mail address: eising2@stolaf.edu

Yuguo Chen
Department of Statistics
University of Illinois at Urbana-Champaign
Champaign, IL 61820
USA
E-mail address: yuguo@illinois.edu