Generating uniformly distributed random networks

Yael Artzy-Randrup* and Lewi Stone†
Biomathematics Unit, Faculty of Life Sciences, Tel Aviv University, Ramat Aviv 69978, Israel
(Received 17 February 2005; revised manuscript received 28 July 2005; published 16 November 2005)

The analysis of real networks taken from the biological, social, and physical sciences often requires a carefully posed statistical null-hypothesis approach. One common method requires comparing real networks to an ensemble of random matrices that satisfy realistic constraints in which each different matrix member is equiprobable. We discuss existing methods for generating uniformly distributed (constrained) random matrices, describe their shortcomings, and present an efficient technique that should have many practical applications.

DOI: 10.1103/PhysRevE.72.056708 PACS number(s): 02.70.−c, 89.75.Hc, 89.75.Fb, 05.45.−a

I. INTRODUCTION

Characterizing the statistical and mathematical properties of complex networks is an exciting multidisciplinary research area having recent significant impact in the fields of biology, social sciences, and physics [1–12,15]. When analyzing real-world systems, it has become common practice to test whether an observed network is different from what one might expect had it been constructed by chance alone—that is, as if all network links were randomly rewired [4–12]. This leads us into the arena of null-hypothesis testing where the statistical features of an observed network are compared to those found in an ensemble of representative random networks. This requires a technique for generating an ensemble of random networks with each ensemble member being equally as likely to occur as any other. However, generating uniformly distributed samples from an ensemble of random networks is a complicated procedure as emphasized by the current controversy in the literature [5–7,10–12,15], and the more common algorithms fail to fulfill this criterion. Here we introduce a method that generates uniformly distributed random samples, is more computationally efficient than existing algorithms, is simple to implement, and should have many practical applications.

A network is a directed graph whose nodes represent a set of "agents," with edges linking those nodes that interact in some specified manner. In the study of biological networks, the nodes might represent genes (/neurons) and the links might represent regulation pathways (/synaptic connections). The degree of any given node is defined as the total number of edges it is attached to. A network of N nodes can be fully defined by a 0-1 binary matrix $\underline{\underline{A}}=[a_{ij}]_{N\times N}$ with $a_{ij}=1$ if a directed link exists from node i to j and $a_{ij}=0$ otherwise. Figure 1(a) makes clear the correspondence between a network and its equivalent matrix representation. The row and column sums of the matrix are given by $r_i=\Sigma_{j=1}^{N}a_{ij}$ and $c_j=\Sigma_{i=1}^{N}a_{ij}$, corresponding to the number of outgoing and ingoing edges of each node in the network, thereby fully defining the degree distribution of all nodes.

The study of 0-1 binary matrices has a long history that is not exclusively confined to networks [16]. In ecology, for example, they are referred to as presence-absence matrices and summarize the appearance of individuals or species at particular habitats. In the field of island biogeography, rows might represent different species, while columns might represent different sites or islands. If species i is present at site j, then $a_{ij}=1$ in the binary presence-absence matrix $\underline{\underline{A}}$; otherwise, $a_{ij}=0$. Presence-absence matrices do not necessarily have an equal number of rows and columns, as do matrices describing a network. Computational and statistical methods for analyzing these matrices in biophysics and biological applications have been a source of great friction over the last three decades [4–7,12–14].

II. NULL-HYPOTHESIS APPROACH

The null-hypothesis approach is based on a comparison between the observed network and an ensemble of networks that are randomly constructed. By comparing the observed data to "all possible worlds" one can deduce whether or not it is significantly unusual and try to identify those features which are responsible for any nonrandomness. In conducting such tests, three ingredients are essential. First, it is important to precisely define the random null hypothesis. Second an algorithm is required for generating random networks that

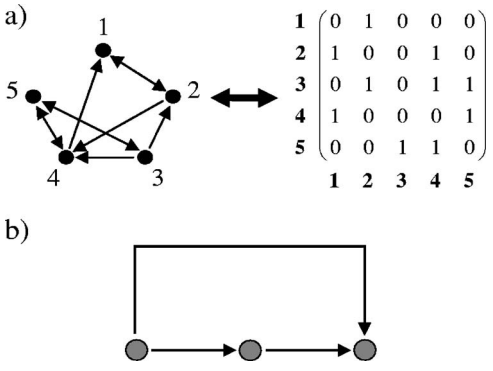

FIG. 1. (a) A typical network consisting of five nodes and seven edges. The binary matrix $\underline{\underline{A}}=[a_{ij}]_{N\times N}$ on the right fully characterizes the network structure. That is, if an edge is connected between nodes i and j, then the matrix entry is set to $a_{ij}=1$; otherwise, $a_{ij}=0$. (b) The feed-forward loop (FFL) is a three-node subgraph with edges connected in the formation shown.

*Electronic address: artzyra@post.tau.ac.il
†Electronic address: lewi@post.tau.ac.il

are truly unbiased or "null." Third, one needs to choose an appropriate test statistic and determine whether the observed score is significantly nonrandom with respect to the distribution of the statistic under the null hypothesis.

When defining the null hypothesis it is necessary to allow for realistic constraints that preserve properties of the observed data—properties that might be considered invariant. One common practice that we adhere to in this paper requires conserving the distribution of both incoming and outgoing edges for all nodes in the network—i.e., the degree distribution. This may be achieved by ensuring that each random matrix inherits the same row and column sums $\underline{r}=\{r_i\}_{i=1}^{N}$ and $\underline{c}=\{c_j\}_{j=1}^{N}$ as the observed system under study. Consider what might arise if the degree distribution of the nodes in the observed network $\underline{\underline{A}}$ was scale free (with a power-law distribution) and the random matrices failed to reflect this degree distribution. In such a case, the null hypothesis might well be rejected for this difference alone, regardless of any unusual characteristics in $\underline{\underline{A}}$ itself.

Thus interest centers on generating independent random samples from the full universe $U(\underline{r},\underline{c})$ of all $|U|$ possible matrices which have the same row and column sums. We note that an explicit formula enumerating $U(\underline{r},\underline{c})$ has been developed [17,18], which might be useful for estimating sample sizes when conducting null-hypothesis tests. Unfortunately, the formula is awkward to work with since for even relatively small matrices $|U(\underline{r},\underline{c})|$ is a large and unwieldly number.

III. SAMPLING BY "SWITCHING"

The switching method [6–9,16] is the simplest and best known technique for generating a random sample of matrices in $U(\underline{r},\underline{c})$. The method takes advantage of "checkerboard" patterns appearing in a matrix:

$$\begin{array}{ccccccccccc} & \vdots & & \vdots & & & & \vdots & & \vdots & \\ \cdots & 1 & \cdots & 0 & \cdots & & \cdots & 0 & \cdots & 1 & \cdots \\ & \vdots & & \vdots & & \Leftrightarrow & & \vdots & & \vdots & \\ \cdots & 0 & \cdots & 1 & \cdots & & \cdots & 1 & \cdots & 0 & \cdots \\ & \vdots & & \vdots & & & & \vdots & & \vdots & \end{array}$$

The checkerboard on the left can be switched to its mirror on the right and vice versa without changing the matrices' row and column sums. Matrices are considered to be "neighbors" if one can be obtained from the other by performing a single switch. Naively, it might be expected that by randomly switching a large number of checkerboard units in the matrix, it is possible to generate a random sample of matrices from $U(\underline{r},\underline{c})$. This is the basis of the popular switching method. As each random switch generates a new neighboring matrix belonging to $U(\underline{r},\underline{c})$, the technique can be formulated as a Markov chain (MC). It has been proven that any matrix in the universe $U(\underline{r},\underline{c})$ can be obtained from any other by some finite number of switches and thus the MC is irreducible [15,16]. Being aperiodic, the MC must eventually converge to a unique stationary distribution [19].

It should be noted that some network studies exclude the possibility of self-loops—i.e., $a_{ii}=0$ for $1\leqslant i\leqslant N$ in the network's corresponding 0-1 matrix. In such cases it is often necessary to also constrain $U(\underline{r},\underline{c})$ to contain only those matrices whose diagonal terms are all $a_{ii}=0$. This class of matrices is not always irreducible, and therefore the above MC formulation, as it stands, might be thought inappropriate. Nevertheless, we can show that even for such cases, the switching method is valid [20].

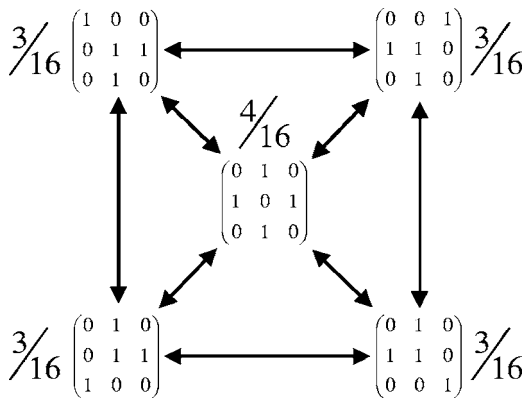

FIG. 2. An example universe of all 3×3 matrices with row sums $\underline{r}=(1,2,1)$ and column sums $\underline{c}=(1,2,1)$ (see also [5,6,22]). This universe has five members which are connected by a network of checkerboard switches. Some members have a higher probability of being switched to, and therefore when sampling this universe randomly via checkerboard switches, the frequencies of the matrices in the sample are not uniform.

As an illustration of the switching method consider a universe $U(\underline{r},\underline{c})$ of all 3×3 matrices with $\underline{r}=(1,2,1)$ and $\underline{c}=(1,2,1)$. This universe has $|U(\underline{r},\underline{c})|=5$ members, which are presented in Fig. 2. An arrow between two matrices indicates that they are neighbors and that it only requires a single switch to transform from one to the other. If the switching is random, then not all matrices will be visited with the same frequency. Hence the sampling is not uniform [5,6,15]. In fact, each matrix will be visited for a time that is proportional to its number of neighbors [5,19]. Thus a random walk through $U(\underline{r},\underline{c})$ will produce matrices with frequencies proportional to their associated number of neighbors. For the case of Fig. 2, four of the matrices have three neighbors and one matrix has four neighbors. Hence the first four matrices will appear with frequency 3/16 and the remaining matrix with frequency 4/16.

It is possible to calculate the unique stationary distribution of the MC produced by the switching method in a more formal fashion [15,19]. Suppose the MC is in the state represented by $\underline{\underline{A}}_i$, a matrix which has a total number of n_i checkerboard units or, equivalently, n_i different neighboring matrices. Let p_{ij} be the probability of moving from matrix $\underline{\underline{A}}_i$ to matrix $\underline{\underline{A}}_j$ in the MC and set

$$p_{ij}=\begin{cases} 1/n_i & \text{if matrix } A_i \text{ and } A_j \text{ are neighbors,} \\ 0 & \text{otherwise.} \end{cases} \tag{1}$$

Take $\underline{\pi}=(\pi_1,\ldots,\pi_{|U|})$ as a probability vector where π_i is the probability of the MC being in the "state" represented by matrix $\underline{\underline{A}}_i$. As a consequence of the ergodic theorem, π_i is proportional to the mean amount of time the MC visits the

state $\underline{\underline{A}}_i$ [19]. With each interchange, the probability distribution of the MC is updated by

$$\underline{\pi}_{t+1} = \underline{\pi}_t \underline{\underline{P}}, \tag{2}$$

where $\underline{\underline{P}}=[p_{ij}]_{N\times N}$ is the transition matrix. For an irreducible and aperiodic MC, the limiting stationary distribution is the probability vector $\underline{\pi}^*=(\pi_1^*,\ldots,\pi_{|U|}^*)$, which fulfills

$$\underline{\pi}^* \underline{\underline{P}} = \underline{\pi}^* \Leftrightarrow \sum_{i=1}^{|U|} \pi_i^* p_{ij} = \pi_j^*. \tag{3}$$

Taking $\alpha_i=n_i/\Sigma_{k=1}^{|U|}n_k$ we find that

$$\sum_{i=1}^{|U|} \alpha_i p_{ij} = \sum_{i=1}^{|U|} \frac{n_i p_{ij}}{\sum_{k=1}^{|U|} n_k} = \frac{1}{\sum_{k=1}^{|U|} n_k} \overbrace{(1+\cdots+1)}^{n_j} = \alpha_j. \tag{4}$$

Hence the stationarity condition is satisfied [Eq. (3)] and

$$\underline{\pi}^* = \underline{\pi}_S^* = (n_1,\ldots,n_{|U|}) \Big/ \sum_{k=1}^{|U|} n_k. \tag{5}$$

That is, each matrix is visited in proportion to the number of checkerboards it contains or equivalently its number of neighbors.

This has the implication that the switching method cannot generate samples of $U(\underline{r},\underline{c})$ that are uniformly distributed. Instead, it is biased—the greater the number of neighbors a matrix has, the more time it will be visited by the MC. For such an MC, the ergodic mean of a chosen statistic f converges to its theoretical mean: $\bar{f}_t \underset{t\to\infty}{\to} \mu_{\underline{\pi}^*_S}$ under the nonuniform distribution $\underline{\pi}_S^*$, where t is the length of the MC.

As a check on this we examine a biological example based on the so-called feed-forward loop (FFL) motif [8,9]. The FFL motif is a particular three-node subgraph [see Fig. 1(b)], named aptly because of its hypothesised role in biological networks. There is large body of work [4,8,9] which aims to test whether the FFL motif is significantly more abundant in biological networks than chance would allow for, in which case the FFL might be viewed as evidence for an evolutionary design principle. Hence, as our test statistic, we let f denote the number of FFL motifs in the matrix under investigation.

Consider the specific universe $U(\underline{r},\underline{c})$ of all 10×10 matrices with $\underline{r}=(3,1,7,2,1,3,7,2,5,9)$ and $\underline{c}=(4,8,1,4,9,3,1,6,3,1)$. We listed all $|U|=2214$ matrices of this universe and calculated the f score for each of these matrices. It was thus possible to calculate the exact theoretical mean ($\mu_{\underline{\pi}^*_S}=58.2$) of f under the stationary distribution $\underline{\pi}_S^*$ given by Eq. (5)—that is, the mean expected to result from implementing the biased switching method. (Note that this differs from the theoretical mean for matrices that are uniformly distributed.) The expected number of FFL's per matrix was found to be $\mu_{\underline{\pi}^*_S}=58.2$ under $\underline{\pi}_S^*$. Figure 3(a) shows this by iterating via the switching method and plotting the mean number of FFL motifs per matrix as a function of sample size. The MC rapidly converges to the mean $\mu_{\underline{\pi}^*_S}=58.2$ FFL's.

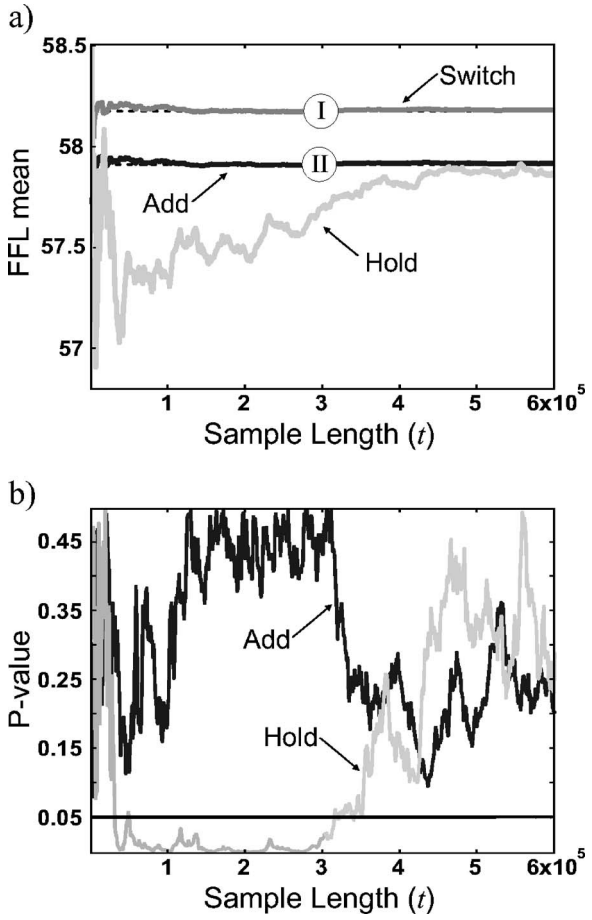

FIG. 3. (a) Mean number of FFL's per matrix generated by the switch, hold, and add methods (marked with arrows), as a function of sample length t (i.e., number of iterations) from $U(\underline{r},\underline{c})$, where $\underline{r}=(3,1,7,2,1,3,7,2,5,9)$ and $\underline{c}=(4,8,1,4,9,3,1,6,3,1)$. All three simulations converge to theoretical predictions (horizontal lines I and II correspond to $\mu_{\underline{\pi}^*_S}$ and $\mu_{\underline{\pi}^*_U}$, respectively). (b) P values of the hold and add methods obtained by a one-sampled t test (see [28]) between the theoretical mean $\mu_{\underline{\pi}^*_U}$ and the ergodic mean $\bar{f}_t$ as function of sample length. The significance level ($\alpha=0.05$) is plotted in black.

IV. SAMPLING BY "SWITCHING AND HOLDING"

The so-called Monte Carlo Markov chain (MCMC) hold method [15,21–23] was developed to sample matrices from $U(\underline{r},\underline{c})$ uniformly and without bias. The method is based on the way in which a checkerboard unit may be randomly selected in a binary matrix. In this scheme a set of two different rows and columns is chosen at random from matrix $\underline{\underline{A}}_i$. If this set falls on a checkerboard unit, a switch is performed, and the newly generated matrix $\underline{\underline{A}}_{i+1}$ is registered as the next state in the MC. However, if the set does not fall on a checkerboard unit, the old matrix $\underline{\underline{A}}_i$ is again registered as the next state in the MC—i.e., $\underline{\underline{A}}_{i+1}=\underline{\underline{A}}_i$—and the MC is said "to hold

on" to this matrix. The process is repeated by finding a new random set of rows and columns and either holding onto the old matrix if this trial fails to fall on a checkerboard configuration or moving onto the appropriate neighboring matrix if a checkerboard is found. This contrasts with the switching method where the MC moves to the next state only when a checkerboard is found. In the hold method every trial, checkerboard or not, leads to the creation of a new state. The sample of matrices so produced is thus comprised of a union of chains of repeats of matrices.

In the hold method the length of each chain is stochastically determined and there is a correlation between this length and the number of neighbors of a matrix. The fewer the neighbors, the smaller is the probability of finding a checkerboard unit, thereby increasing the chance the MC holds onto the matrix and leading to a longer chain of repeats. The outcome is a negative bias for matrices with many neighbors (favored by the switching method) and a positive bias for matrices with fewer neighbors. The scheme converges to a uniform distribution of frequencies of the different matrices in $U(\underline{r},\underline{c})$. To see this formally, define the MCMC transition matrix $\underline{\underline{P}}$ as:

$$p_{ij} = \begin{cases} 1/Q_N & \text{matrices } A_i \text{ and } A_j \text{ are neighbors,} \\ 1 - n_i/Q_N & \text{for } i = j, \\ 0 & \text{otherwise,} \end{cases} \tag{6}$$

where $Q_N=[N(N-1)/2]^2$ is the number of possible sets of pairs of rows and columns one can choose in an $N\times N$ matrix [24]. These relations satisfy what are referred to as the detailed balance equations

$$\{\alpha_i p_{ij} = \alpha_j p_{ji}\}_{i,j\in U}, \tag{7}$$

where $\alpha_i=1/|U|$ is the target uniform distribution. Therefore,

$$\sum_{j=1}^{|U|} \alpha_j p_{ji} = \sum_{j=1}^{|U|} \alpha_i p_{ij} = \alpha_i \sum_{j=1}^{|U|} p_{ij} = \alpha_i \tag{8}$$

and thus

$$\underline{\pi}^* = \underline{\pi}^*_U = (1, \ldots, 1)/|U| \tag{9}$$

is the limiting stationary distribution [Eq. (3)]. Hence the hold method leads to a stationary state that is uniformly distributed. This is demonstrated in Fig. 3(a) where the average number of FFL motifs [same $U(\underline{r},\underline{c})$ as in the previous example] is plotted as a function of sample size for random matrices generated by the hold method. The MC generated by the hold method converges to the theoretical mean $\mu_{\underline{\pi}^*_U}=57.9$ as calculated exactly for uniformly distributed random matrices $\underline{\pi}^*_U$.

A significant drawback of the hold method arises because the probability of repeating or holding onto a matrix is given by p_{ii} [as defined in Eq. (6)] and is in general very large. An analysis of a wide range of random matrices of different densities shows that typically $p_{ii}>0.87$, meaning that in general more than 87% of the trial swaps fail to land on a checkerboard unit. This makes the hold method extremely inefficient and leads to long and redundant chains of copies. As a result, the number of distinct matrices is greatly reduced, as is the diversity of the random sample of matrices. Thus for the hold method to give a reasonable estimate of the universe of which it is being drawn from, the sample must be very large, as can be seen from Fig. 3(a).

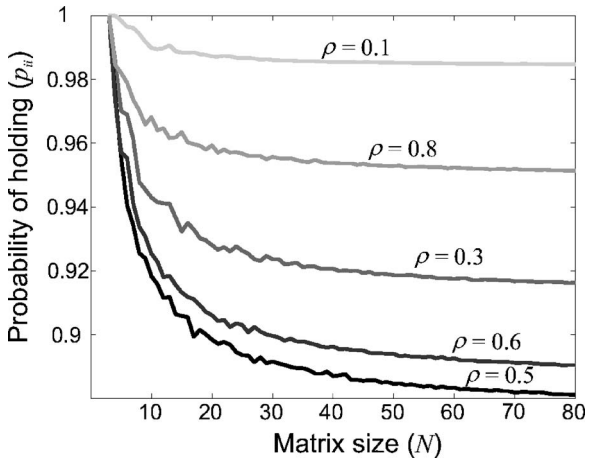

FIG. 4. We analyzed $N\times N$ matrices for a range of sizes from $N=5$ to $N=100$. Matrices were filled randomly with 1's at different densities, denoted as $\rho=$(number of ones)$/(N\times N)$. The mean probability of holding was calculated through simulations and tended to decrease with matrix size N but was always large with $p_{ii}>0.87$. Notice that for matrices with density ρ and for matrices with density $1-\rho$ the mean probability of holding is equal, and so for matrices with $\rho=0.5$ the mean probability of being held on to is lowest. For $\rho=0.5$ the minimum probability of being held is $p_{ii}=0.75$ and the maximum is $p_{ii}=1$, with the mean probability being $p_{ii}>0.87$.

It is possible to quantify this further. Based on a large-scale analysis of simulations we conjecture that the maximum number of neighbors an $N\times N$ matrix has is $n_{max}=(N/2)^4$, causing $p_{ii}=1-n_{max}/Q_N\to 0.75$ for large N's (i.e., the maximum number of checkerboards a 12×12 matrix can hold is 1296, and thus $p_{ii}=0.7025$). Since the majority of $N\times N$ matrices have much fewer neighbors than n_{max}, the probability of holding on to them in the MC chain is $p_{ii}\gg 0.75$, and for some cases the probability can be close to 1 (see Fig. 4). It follows that the chains of repeats tend to be very long and the variety of different independent matrices sampled is low. As a general example, consider a fictitious population in which each item, once sampled, has a 0.9 probability of being resampled. A sample set assembled from 100 draws would enclose on average only $\sim$11 distinct items rather than 100 independent samples. The rest of this sample set would in practice contain repeats of these $\sim$11 items.

V. SAMPLING BY "SWITCHING AND ADDING"

Here we propose the add method for uniformly generating samples from $U(\underline{r},\underline{c})$. The method takes advantage of computational techniques to locate and list all n_i checkerboards of each new matrix $\underline{\underline{A}}_i$ in the MC. This obviates going through the inefficient search process of randomly stumbling upon sets of rows and columns to locate a checkerboard unit. With the checkerboards located *ab initio*, the probabilities p_{ij}

of the transition matrix in Eq. (6) can be assigned directly and used to make the decision of holding on to the same matrix or advancing to a new one. Each time a matrix $\underline{\underline{A}}_i$ is generated by the MC it has a probability of $p_{ii}=1-n_i/Q_N$ for being reregistered, thus generating a chain of repeats. The chain is composed of a series of failures (with probability p_{ii}) corresponding to "holds" and terminates with a single success that corresponds to finding a switch. It thus has a geometric distribution with expected length

$$L_i = 1 + \sum_{j=1}^{\infty} j(1-p_{ii})p_{ii}^j = 1 + p_{ii}/(1-p_{ii}) = Q_N/n_i. \quad (10)$$

The add method takes into account long-term averages by directly representing every matrix generated by the MC for a time period (in terms of iterations) that is proportional to its expected hold time L_i. That is, when the MC generates matrix $\underline{\underline{A}}_i$, L_i copies of this matrix are immediately added to the MC sample. In the long run, the add method must represent matrices in the same proportions as the hold method. In practice, for each matrix $\underline{\underline{A}}_i$ generated by the MC, the matching number of neighbors n_i is recorded. Once the MC ends its course, these recorded values of neighbors may be retroactively used to determine how many times each matrix should hypothetically be held on to.

Seen in another way, according to Eq. (10) each matrix should be weighted by a factor that is inversely proportional to the number of checkerboard units it contains. The result agrees with Zaman and Simberloff [5]. This is an intuitively pleasing result since in the switching method Eq. (5) implies that matrices are visited in proportion to their number of checkerboards, but the weighting of the add scheme completely compensates for this effect yielding uniformity in distribution.

A potential concern in implementing the add method is that the values of $L_i=Q_N/n_i$ in Eq. (10) are generally fractional. If necessary, the L_i can always be transformed to integer values by multiplication with a common number C (e.g., the lowest common multiplier of the n_i), a procedure that conserves their relative ratios. Conversely, this same reasoning reveals why it is permissible to use fractional (rather than integral) chain lengths $L_i=Q_N/n_i$. For example, when estimating the mean of a statistic f_i from a sample with the add method, we use the formula

$$\bar{f}_t = \frac{\sum_{i=1}^{t} f_i CL_i}{\sum_{i=1}^{t} CL_i} = \frac{\sum_{i=1}^{t} \frac{f_i}{n_i}}{\sum_{i=1}^{t} \frac{1}{n_i}}, \quad (11)$$

where C is some common multiplier of all the n_i, such that CL_i is an integer for all i. As the Q_N and C cancel out, the sample mean $\bar{f}_t$ reduces to the familiar "weighted mean," where each weight correlates to the probability of being sampled, $w_i=1/n_i$.

The weighting scheme of the add method may be easily understood by returning back to the simple example in Fig. 2 where matrices $U(\underline{r},\underline{c})$ contains only five distinct matrices. Their relative frequencies need to be weighted in inverse proportion to their respective number of neighbors. The first four matrices will thus have relative frequencies $1/3\times3/16$ and the fifth matrix will have relative frequency $1/4\times4/16$. That is, after the weighting, all matrices are equiprobable.

VI. COMPARING THE SWITCHING, HOLD, AND ADD METHODS

Figure 3(a) provides a comparison of the switching, hold, and add Methods again for the example universe $U(\underline{r},\underline{c})$ of all 10×10 matrices with $\underline{r}=(3,1,7,2,1,3,7,2,5,9)$ and $\underline{c}=(4,8,1,4,9,3,1,6,3,1)$. With respect to the FFL motif statistic f, it is clear that the add method converges far more rapidly to the exact mean $\mu_{\underline{\underline{\pi}}_U^*}=57.9$ than the hold method, while as we have seen the switching method, being biased, converges to a different mean altogether. A one-sample t test (see [28]) was conducted between the exact mean $\mu_{\underline{\underline{\pi}}_U^*}$ and sample mean $\bar{f}_t$ for both the hold method and the add method as the MC simulation in Fig. 3(a) progressed, giving a p value as a function of sample size [Fig. 3(b)]. In contrast to the add method, the sample mean generated by the hold method stays significantly different from the exact mean even for large sample sizes ($>10^5$). As some of the key network studies in the literature have relied on the hold method with sample sizes of ≈1000 matrices, this may be a cause for concern. Figure 3(b) makes clear that the sample size of these studies might be underestimated by several orders of magnitude.

VII. RUN TIME

The superiority of the add method is due to several reasons. Recall that in the hold method the main motivation for randomly picking rows and columns is not for finding possible checkerboard units (there are more efficient methods), but for determining through trial and error the length of the chains of repeats. In contrast, with the add method each matrix in this scheme is "held" for a period of time that is calculated instantaneously and deterministically, rather than by repeatedly "flipping a coin." By analytically calculating these values, not only does the add method spare unnecessary computational loops, but it also delivers precise values to act as a weights needed to counteract the natural bias induced by switching. This helps to speed up convergence to stationarity.

It has been brought to our attention that the add method belongs to a class of event-induced algorithms, pioneered by Bortz, Kalos, and Lebowitz [25] and has been used in different areas of computational physics [26,27]. For example, when simulating the low-temperature relaxation of spin glasses, instead of having an algorithm iterate through many rejections, the waiting time method [27] calculates an expected average waiting time. The algorithm then jumps immediately to its next state at the appropriate moment without iterations. By saving extensive computations, this approach is far more efficient.

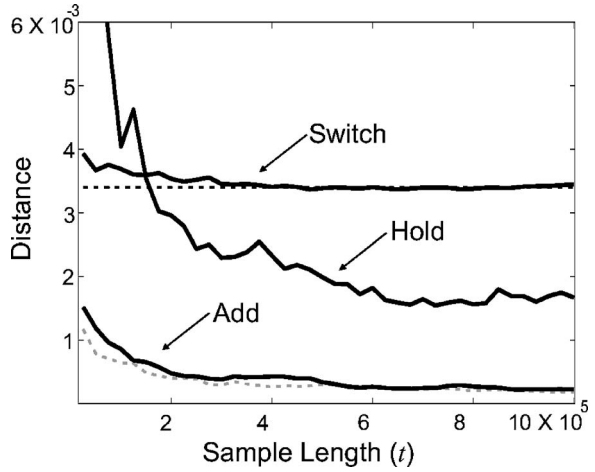

FIG. 5. The distance $\|\underline{\pi}^*_U - \underline{d}^{(t)}\|$ [see Eq. (12)] is used as an index of convergence to stationarity and plotted as a function of sample length t for the three methods based on the universe $U(\underline{r},\underline{c})$ of 7×7 matrices (details in text). The switch method converges to theoretically predicted distance (upper black dashed line). The add method converges to zero in a manner similar to ball-urn sampling experiment (lower gray dashed line; see text). The hold method converges towards zero as well, but at a much slower pace.

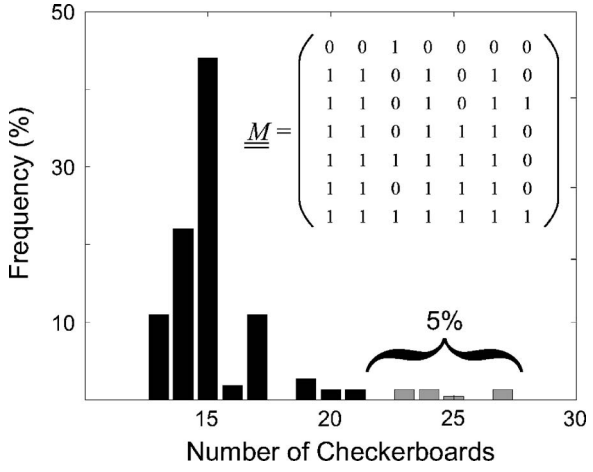

FIG. 6. Frequency histogram of the distribution of the checkerboard score (total number of checkerboards) in all matrices of U. Matrix $\underline{\underline{M}}$ has 23 checkerboards and is thus considered to be unusual because it lies in the 5% significant region (in gray).

The rapid convergence of the add method is even more transparent in Fig. 5, plotted for the universe $U(\underline{r},\underline{c})$ with $\underline{r}=(1,4,5,5,6,5,7)$ and $\underline{c}=(6,6,3,6,4,6,2)$ which consists of $|U|=218$ matrices. In this figure we plot the distance between observed (normalized) frequencies $\underline{d}^{(t)}=(d_1^{(t)},\ldots,d_{|U|}^{(t)})$ of matrices generated by the MC after t iterations and the true stationary distribution $\underline{\pi}^*=\underline{\pi}^*_U=(1,\ldots,1)/|U|$. We define the distance as

$$\|\underline{\pi}^*_U - \underline{d}^{(t)}\| = \sup_{A_i\in U} |\pi_{U_i} - d_i^{(t)}| \tag{12}$$

and plot the distance as a function of the number of matrices generated by the MC. The three different sampling methods were used to generate the vector $\underline{d}^{(t)}$. Convergence to stationary frequencies requires that $\|\underline{\pi}^*_U - \underline{d}^{(t)}\| \underset{t\to\infty}{\to} 0$. In Fig. 5, one immediately sees that the switch method fails to converge to a uniform distribution, while the add method converges far more rapidly than the hold method.

In order to understand the add method's convergence rate better we compared it to t balls being dropped randomly into a set of $|U|=218$ urns with equal probability. Let $\underline{d}^{(t)}=(d_{A_1},\ldots,d_{A_{|U|}})$ be the observed (normalized) frequencies of the balls in the urns. Figure 5 plots the distance $\|\underline{\pi}^*_U - \underline{d}^{(t)}\|$ as a function of t and makes clear that the balls converge to a uniform distribution at what appears to be the same rate as the add method. The comparison shows that the convergence of the add method is set in the main by the sampling process itself.

VIII. IMPLEMENTING THE NULL-HYPOTHESIS TEST

As an application of the add method consider the matrix $\underline{\underline{M}}=[m_{ij}]_{N\times N}$ shown in Fig. 6 belonging to the universe $U((1,4,5,5,6,5,7),(6,6,3,6,4,6,2))$ of 7×7 matrices. This matrix describes a group of seven scientists interested in seven different topics of research, such that each row in the matrix represents a scientist and each column represents a topic. If a scientist i is interested in topic j, then $m_{ij}=1$; otherwise, $m_{ij}=0$. Assuming that some topics attract wider interest than others and that some scientists have more diverse interests, we raise the following question: Is the distribution of interests between scientists a matter of chance, or do these particular seven scientists have some nonrandom pattern of interest? For example, there might be a tendency for scientists to be more drawn towards certain topics or to avoid topics their colleagues are already working on. To the naked eye, this matrix does not appear unusual, and it was thus of interest to subject the matrix to the random null-hypothesis test. We compared the matrix to the entire universe of all possible matrices sharing these constraints [i.e., the universe $U((1,4,5,5,6,5,7),(6,6,3,6,4,6,2))$]. As a test statistics, we counted the number of times a scientist i_1 was interested in topic j_1 while another scientist i_2 was interested in topic j_2, such that i_1 was not interested in j_2, and i_2 was not interested in j_1. This corresponds to the number of checkerboard patterns between all scientist i_1 and i_2. The total number of such checkerboards in matrix $\underline{\underline{M}}$ was found to be $n=23$. The distribution of checkerboard scores found in the universe $U(r,c)$ as sampled uniformly by the add method is shown as a frequency histogram in Fig. 6. One sees that the number of checkerboards in $\underline{\underline{M}}$ is unusual and significantly overrepresented ($p=0.04$), lying in the 5% critical region of the frequency histogram. Thus the interests of the scientists is indeed nonrandom and there is an excess amount of exclusion patterns whereby pairs of scientists tend to avoid working on the same topic. This result may be reproduced by using the exact distribution of checkerboards found from listing all $|U|=218$ matrices. However, if the same test is carried out using the nonuniform switching method to generate a null model, a contrary result is obtained and the number of checkerboards in the above matrix is not significant

($p=0.07$). Thus the add method gives the correct interpretation while the switching method fails.

Finally, we note that we are able to generalize this method so that it is also applicable for networks that lack self-loops ($a_{ii}=0$ for all i), as will shortly be reported elsewhere.

ACKNOWLEDGMENT

We gratefully acknowledge the support of the James S. McDonnell Foundation, and thank Professor P. Sibani for helpful suggestions.

[1] D. J. Watts and S. H. Strogatz, Nature (London) **393**, 440 (1988).
[2] A. L. Barabasi and R. Albert, Science **286**, 509 (1999).
[3] S. Maslov and K. Sneppen, Science **296**, 910 (2002).
[4] Y. Artzy-Randrup, S. Fleishman, N. BenTal, and L. Stone, Science **305**, 1107c (2004).
[5] A. Zaman and D. Simberloff, Environ. Ecol. Stat. **4**, 405 (2002).
[6] L. Stone and A. Roberts, Oecologia **85**, 74 (1990).
[7] A. Roberts and L. Stone, Oecologia **83**, 560 (1990).
[8] R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon, Science **298**, 824 (2002).
[9] R. Milo, S. Itzkovitz, N. Kashtan, R. Levitt, S. Shen-Orr, I. Ayzenshtat, M. Sheffer, and U. Alon, Science **303**, 1538 (2004).
[10] O. D. King, Phys. Rev. E **70**, 058101 (2004).
[11] S. Itzkovitz, R. Milo, N. Kashtan, M. E. J. Newman, and U. Alon, Phys. Rev. E **70**, 058102 (2004).
[12] N. J. Gotelli and G. R. Graves, *Null Models in Ecology* (Smithsonian Institution Press, Washington, DC, 1996).
[13] N. J. Gotelli and D. J. McCabe, Ecology **83**, 2091 (2002).
[14] B. F. J. Manly and J. G. Sanderson, Ecology **83**, 580 (2002).
[15] A. R. Rao, R. Jana, and S. Bandyopadhya, Sankhya, Ser. A **58**, 225 (1996).
[16] H. J. Ryser, *Combinatorial Mathematics* (The Mathematical Association of America, Buffalo, NY, 1963).
[17] B. R. Perez-Salvador, S. de-los-Cobos-Silva, M. A. Gutierrez-Andrade, and A. Torres-Chazaro, Discrete Math. **256**, 361 (2002).
[18] B. Y. Wang and F. Zhang, Discrete Math. **187**, 211 (1998).
[19] S. M. Ross, *Stochastic Processes* (Wiley, New York, 1996).
[20] See http://www.tau.ac.il/lifesci/departments/zoology/members/stone/stone.html
[21] R. Milo, N. Kashtan, S. Itzkovitz, M. E. J. Newman, and U. Alon, e-print cond-mat/0312028.
[22] I. Miklos and J. Podani, Ecology **85**, 86 (2004).
[23] G. W. Cobb, www.mtholyoke.edu/courses/gcobb/stat344/book.html
[24] For matrices representing networks with no self-loops, $Q_N=N(N-1)(N-2)(N-3)/4$ is the number of "legal" pairs of rows and columns one can choose from.
[25] A. B. Bortz, M. H. Kalos, and J. L. Lebowitz, J. Comput. Phys. **17**, 10 (1975).
[26] D. T. Gillespie, J. Comput. Phys. **22**, 403 (1976).
[27] J. Dall and P. Sibani, Comput. Phys. Commun. **141**, 260 (2001).
[28] Sequential samples were separated from each other by 1000 switches to prevent dependency. For the hold method the argodic mean is $\bar{f}_t=\Sigma_{i=1}^{C_t}H_i f_i/t$ and the variance is $\bar{s}_t^2=[C_t/(C_t-1)][\overline{(f^2)}_t-(\bar{f}_t)^2]$, where C_t is the number of chains in a sample of length t and H_i is the length of each such chain. For the add method, $\bar{f}_t=(\Sigma_{i=1}^{t}L_i f_i)/(\Sigma_{i=1}^{t}L_i)$ and $\bar{s}_t^2=[t/(t-1)][\overline{(f^2)}_t-(\bar{f}_t)^2]$, such that for the sampled matrix at state i, L_i is the expected chain length [Eq. (10)].