# The Number of Attractors in Kauffman Networks

Björn Samuelsson* and Carl Troein†
*Complex Systems Division, Department of Theoretical Physics*
*Lund University, Sölvegatan 14A, S-223 62 Lund, Sweden*
(Dated: January 17, 2003)

The Kauffman model describes a particularly simple class of random Boolean networks. Despite the simplicity of the model, it exhibits complex behavior and has been suggested as a model for real world network problems. We introduce a novel approach to analyzing attractors in random Boolean networks, and applying it to Kauffman networks we prove that the average number of attractors grows faster than any power law with system size.

PACS numbers: 89.75.Hc, 02.70.Uu

## INTRODUCTION

We are increasingly often faced with the problem of modeling complex systems of interacting entities, such as social and economic networks, computers on the Internet, or protein interactions in living cells. Some properties of such systems can be modeled by Boolean networks. The appeal of these networks lies in the finite (and small) number of states each node can be in, and the ease with which we can handle the networks in a computer.

A deterministic Boolean network has a finite number of states. Each state maps to one state, possibly itself. Thus, every network has at least one cycle or fixed point, and every trajectory will lead to such an attractor. The behavior of attractors in Boolean networks has been investigated extensively, see e.g. [1–6]. For a recent review, see [7].

A general problem when dealing with a system is finding the set of attractors. For Boolean networks with more than a handful of nodes, state space is too vast to be searched exhaustively. In some cases, a majority of the attractor basins are small and very hard to find by random sampling. One such case is the Kauffman model [8]. Based on experience with random samplings, it has been commonly believed that the number of attractors in that model grows like the square root of system size. Lately this has been brought into question [9–12]. Using an analytic approach, we are able to prove that the number attractors grows faster than any power law. The approach is based on general probabilistic reasoning that may also be applied to other systems than Boolean networks. The derivations in the analysis section are just one application of the approach. We have attempted to focus on the method and our results, while keeping the mathematical details at a minimum level needed for reproducibility using standard mathematical methods.

In 1969 Kauffman introduced a type of Boolean networks as a model for gene regulation [8]. These networks are known as $N$-$K$ models, since each of the $N$ nodes has a fixed number of inputs $K$. A Kauffman network is synchronously updated, and the state (0 or 1) of any node at time step $t$ is some function of the state of its input nodes at the previous time step. An assignment of states to all nodes is referred to as a *configuration*. When a single network, a *realization*, is created, the choice of input nodes and update functions is random, although the update functions are not necessarily drawn from a flat distribution. This reflects a null hypothesis as good or bad as any, if we have no prior knowledge of the details of the networks we wish to model.

In this letter, we will first present our approach, and then apply it to Kauffman's original model, in which there are 2 inputs per node and the same probability for all of the 16 possible update rules. These 16 rules are the Boolean operators of two or fewer variables: AND, OR, TRUE, etc. This particular $N$-$K$ model falls on the *critical line*, where the network dynamics is neither ordered nor chaotic [9, 13, 14].

## APPROACH

Our basic idea is to focus on the problem of finding the number of cycles of a given length $L$ in networks of size $N$. As we will see, the discreteness of time makes it convenient to handle cycles as higher-dimensional fixed point problems. Then it is possible to do the probabilistic averaging in a way which is suitable for an analytic treatment. This idea may also be expanded to applications with continuous time, rendering more complicated but also more powerful methods.

We use four key assumptions: (I) the rules are chosen independently of each other and of $N$, (II) the input nodes are independently and uniformly chosen from all $N$ nodes, (III) the dynamics is dominated by stable nodes, and (IV) the distribution of rules is invariant due to inversion of any set of inputs. (IV) means e.g. that the fraction of AND and NOR gates are the same whereas the fraction AND and NAND gates may differ. (IV) is presumably not necessary, but simplifies the calculations drastically. (III) is expected to be valid for any non-chaotic network obeying (I) and (II) [15]. Note that (I) does not mean that the number of inputs must be the same for every rule. We could write a general treatment of all models obeying

(I) – (IV), but for simplicity we focus on the Kauffman model.

We will henceforth use $\langle C_L \rangle_N$ to denote the expectation value of the number of $L$-cycles over all networks of size $N$, with $L = 1$ referring to fixed points. The average number of fixed points, $\langle C_1 \rangle_N$, is particularly simple to calculate. For a random choice of rules, (I) and (IV) imply that the output state of the net is independent of the input state. Hence, the input and output states will on average coincide once on enumeration of all input states. This means that $\langle C_1 \rangle_N = 1$.

The problem of finding other $L$-cycles can be transformed to a fixed point problem. Assume that a Boolean network performs an $L$-cycle. Then each node performs one of $2^L$ possible time series of output values. Consider what a rule does when it is subjected to such time series on the inputs. It performs some boolean operation, but it also delays the output, giving a one-step difference in phase for the output time series. If we view each time series as a state, we have a fixed point problem. $L \langle C_L \rangle_N$ is then the average number of input states (time series), for the whole network, such that the output is the same as the input.

To take advantage of assumption (IV), we introduce the notion of $L$-cycle patterns. An $L$-cycle pattern is $s$ and $s$ inverted, where $s$ is a time series with period $L$. Let $\mathbf{Q}$ denote a choice of $L$-cycle patterns for the net, and let $P(\mathbf{Q})$ denote the probability that the output of the net is $\mathbf{Q}$. Using the same line of reasoning as for fixed points, we conclude that (I) and (IV) yield

$$\langle C_L \rangle_N = \frac{1}{L} \sum_{\mathbf{Q} \in \mathcal{Q}_L^N} P(\mathbf{Q}) \tag{1}$$

where $\mathcal{Q}_L^N$ is the set of proper $L$-cycles of an $N$-node net. A proper $L$-cycle has no period shorter than $L$.

## ANALYTIC CALCULATIONS

Assumption (II) implies that $P(\mathbf{Q})$ is invariant under permutations of the nodes. Let $\mathbf{n} = (n_0, \ldots, n_{m-1})$ denote the number of nodes expressing each of the $m = 2^{L-1}$ patterns. For $n_j$, we refer to $j$ as the pattern index. For convenience, let the constant pattern have index 0. Then

$$\langle C_L \rangle_N = \frac{1}{L} \sum_{\mathbf{n} \in \mathcal{P}_L^N} \binom{N}{\mathbf{n}} P(\mathbf{Q}) \tag{2}$$

where $\binom{N}{\mathbf{n}}$ denotes the multinomial $N!/(n_0! \cdots n_{m-1}!)$ and $\mathcal{P}_L^N$ is the set of partitions $\mathbf{n}$ of $N$ such that $\mathbf{Q} \in \mathcal{Q}_L^N$. That is, $\mathbf{n}$ represents a proper $L$-cycle.

What we have this far is merely a division of the probability of an $L$-cycle into probabilities of different flavors of $L$-cycles. Now we assume that each node has 2 inputs. Then, we get a simple expression for $P(\mathbf{Q})$ that inserted into Eq. (2) yields

$$\langle C_L \rangle_N = \frac{1}{L} \sum_{\mathbf{n} \in \mathcal{P}_L^N} \binom{N}{\mathbf{n}} \times \prod_{\substack{0 \le j < m \\ n_j \neq 0}} \left( \sum_{0 \le l_1, l_2 < m} \frac{n_{l_1} n_{l_2}}{N^2} (P_L)^j_{l_1 l_2} \right)^{n_j} \tag{3}$$

where $(P_L)^j_{l_1 l_2}$ denotes the probability that the output pattern of a random 2-input rule has index $j$, given that the input patterns have the indices $l_1$ and $l_2$ respectively. Note that Eq. (3) is an exact expression for the average number of proper $L$-cycles in an $N$-node random Boolean network which satisfies the assumptions (I), (II), (IV), and where each node has 2 inputs.

From now on, we only consider the Kauffman model, meaning that we also restrict the distribution of rules to be uniform. It is instructive to explore some properties of $(P_L)^j_{l_1 l_2}$; these will also be needed in the following calculations. We see that

$$(P_L)^0_{00} = 1, \quad (P_L)^0_{l_1 0} = \tfrac{1}{2} \quad \text{and} \quad (P_L)^0_{l_1 l_2} \ge \tfrac{1}{8} \tag{4}$$

for $1 \le l_1, l_2 < m$. Further, we note that for a given $j \neq 0$, $(P_L)^j_{l_1 0}$ has a non-zero value for exactly one $l_1 \in \{1, \ldots, m-1\}$. Let $\phi_L(j)$ denote that value of $l_1$. We can see $\phi_L$ as a function that rotates an $L$-cycle pattern one step backwards in time. With this in mind we define $\phi_L(0) = 0$. Now, we can write

$$(P_L)^j_{l_1 0} = \tfrac{1}{2} \delta_{l_1 \phi_L(j)} \tag{5}$$

for $1 \le j < m$. ($\delta$ is the Kronecker delta.)

We can view $\phi_L$ as a permutation on the set $\{0, \ldots, m-1\}$. Thus, we divide this index space into permutation cycles which are sets of the type $\{j, \phi_L(j), \phi_L \circ \phi_L(j), \ldots\}$. We refer to these permutation cycles as invariant sets of $L$-cycles. Let $\rho_L^0, \ldots, \rho_L^{H_L - 1}$ denote the invariant sets of $L$-cycles, where $H_L$ is the number of such sets. For convenience, let $\rho_0$ be the invariant set $\{0\}$. If two $L$-cycle patterns belong to the same invariant set, they can be seen as the same pattern except for a difference in phase.

We want to find the behavior of $\langle C_L \rangle_N$, for large $N$, by approximating Eq. (3) with an integral. To do this, we use Stirling's formula $n! \approx (n/e)^n \sqrt{2\pi n}$ while noting that the boundary points where $n_j = 0$ for some $j$ can be ignored in the integral approximation. Let $x_j = n_j / N$ for $j = 0, \ldots, m-1$ and integrate over $\mathbf{x} = (x_1, \ldots, x_{m-1})$. $x_0$ is implicitly set to $x_0 = 1 - \sum_{j=1}^{m-1} x_j$. We get

$$\langle C_L \rangle_N \approx \frac{1}{L} \left( \frac{N}{2\pi} \right)^{(m-1)/2} \int_{0 < x_0, \ldots, x_{m-1}} d\mathbf{x} \, \frac{e^{N f_L(\mathbf{x})}}{\prod_{j=0}^{m-1} \sqrt{x_j}} \tag{6}$$

where

$$f_L(\mathbf{x}) = \sum_{j=0}^{m-1} x_j \ln\left(\frac{1}{x_j}\sum_{0\le l_1,l_2<m} x_{l_1}x_{l_2}(P_L)^j_{l_1l_2}\right) . \quad (7)$$

Eq. (7) can be seen as an average $\langle lnX\rangle$, where $X$ is the expression inside the parentheses. Hence, the concavity of $x \to \ln x$ gives $f_L(\mathbf{x}) = \langle lnX\rangle \le ln\langle X\rangle = 0$ with equality if and only if

$$x_j = \sum_{0\le l_1,l_2<m} x_{l_1}x_{l_2}(P_L)^j_{l_1l_2} \quad (8)$$

for all $j = 0,\ldots,m-1$.

Note that Eq. (8) can be interpreted as a mean-field equation of the model. Using Eq. (8) for $j = 0$ and Eq. (4) we see that $f_L(\mathbf{x})$ comes arbitrarily close to zero only in the vicinity of $\mathbf{x} = 0$, and for large $N$, the relevant contributions to the integral in Eq. (6) come from this region. Thus, the dynamics of the net is dominated by stable nodes, in agreement with [11, 15]. This means that assumption (III) is satisfied by the Kauffman model. Using Eqs. (4) and (5), a Taylor-expansion of $f_L(\epsilon\mathbf{x})$ yields

$$\begin{aligned} f_L(\epsilon\mathbf{x}) &= \epsilon\sum_{j=1}^{m-1} x_j \ln\frac{x_{\phi(j)}}{x_j} + \epsilon^2\sum_{j=0}^{m-1} x_j\frac{\mathbf{x}\cdot A^j_L\mathbf{x}}{x_{\phi(j)}} \\ &\quad -\frac{\epsilon^3}{2}\sum_{j=1}^{m-1} x_j\left(\frac{\mathbf{x}\cdot A^j_L\mathbf{x}}{x_{\phi(j)}}\right)^2 + O\left(\epsilon^4\right) \end{aligned} \quad (9)$$

where $(A^j_L)_{l_1l_2} = (P_L)^j_{l_1l_2} - \frac{1}{2}(\delta_{l_1\phi(j)} + \delta_{l_2\phi(j)})$.

The first order term of Eq. (9) has 0 as its maximum and reaches this value if and only if $x_{\phi_L(j)} = x_j$ for all $j = 1,\ldots,m-1$. The second order term is zero at these points, while the third order term is less than zero for all $\mathbf{x} \neq 0$. Hence, the first and third order terms are governing the behavior for large $N$. Using the saddle-point approximation, we reduce the integration space to the space where the first and second order terms are 0. Let $z_h = N^{1/3}\sum_{j\in\rho^h_L} x_j$ for $h = 1,\ldots,H_L-1$ and let $(P'_L)^h_{k_1k_2}$ denote the probability that the output pattern of a random rule belongs to $\rho^h_L$, given that the input patterns are randomly chosen from $\rho^{k_1}_L$ and $\rho^{k_2}_L$ respectively. Thus, we approximate Eq. (6) for large $N$ as

$$\langle C_L\rangle_N \approx \alpha_L\beta_L N^{\gamma_L} \quad (10)$$

where

$$\alpha_L = \left(L\prod_{h=1}^{H_L-1}\left|\rho^h_L\right|\right)^{-1}\left(\frac{1}{2\pi}\right)^{(H_L-1)/2} \quad (11)$$

$$\beta_L = \int_{0<z_1,\ldots,z_{H_L-1}} d\mathbf{z}\,\frac{\exp\left(-\frac{1}{2}\sum_{h=1}^{H_L-1}\frac{1}{z_h}\left(\mathbf{z}\cdot B^h_L\mathbf{z}\right)^2\right)}{\prod_{h=1}^{H_L-1}\sqrt{z_h}} \quad (12)$$

$$\gamma_L = \frac{H_L-1}{3} \quad (13)$$

and $(B^h_L)_{k_1k_2} = (P'_L)^h_{k_1k_2} - \frac{1}{2}(\delta_{h_1h} + \delta_{h_2h})$. ($|\rho|$ denotes the number of elements of the set $\rho$.)

$H_L$ grows rapidly with $L$. The number of elements in an invariant set of $L$-cycle patterns is a divisor of $L$. If an invariant set consists of only one pattern, it is either the constant pattern, or the pattern with alternating zeros and ones. The latter is only possible if $L$ is even. Thus, $H_L - 1 \ge (2^{L-1}-1)/L$, with equality if $L$ is a prime number $> 2$. Applying this conclusion to Eqs. (13) and (10), we see that for any power law $N^\xi$, we can choose an $L$ such that $\langle C_L\rangle_N$ grows faster than $N^\xi$.

## NUMERICAL RESULTS

We have written a set of programs to compute the number of $L$-cycles in Kauffman networks, Eq. (3), both by complete enumeration and using Monte Carlo methods, and tested them against complete enumeration of the networks with $N \le 4$. The results for $2 \le L \le 6$ are shown in Fig. 1, along with the corresponding asymptotes. The asymptotes were obtained by Monte Carlo integration of Eq. (12). For low $N$, $\langle C_L\rangle_N$ is dominated by the boundary points neglected in Eq. (6), and its qualitative behavior is not obvious in that region.

A straighforward way to count attractors in a network is to simulate trajectories from random configurations and count distinct target attractors. As has been pointed out in [11], this gives a biased estimate. Simulations in [8] with up to 200 trajectories per network indicated $\sqrt{N}$ scaling with system size, whereas [11] reported linear

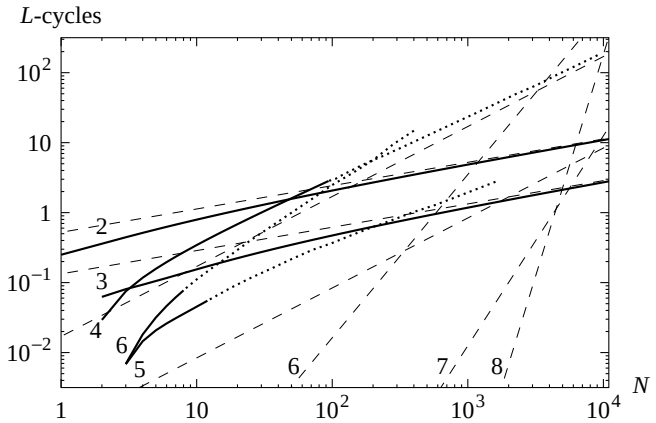


FIG. 1: The number of $L$-cycles as functions of the network size for $2 \le L \le 6$. The numbers in the figure indicate $L$. Dotted lines are used for values obtained by Monte Carlo summation, with errors comparable to the line width. The asymptotes for $N \le 8$ have been included as dashed lines. Their slopes are $\gamma_2 = \gamma_3 = \frac{1}{3}$, $\gamma_4 = \gamma_5 = 1$, $\gamma_6 = \frac{7}{3}$, $\gamma_7 = 3$, and $\gamma_8 = \frac{19}{3}$.

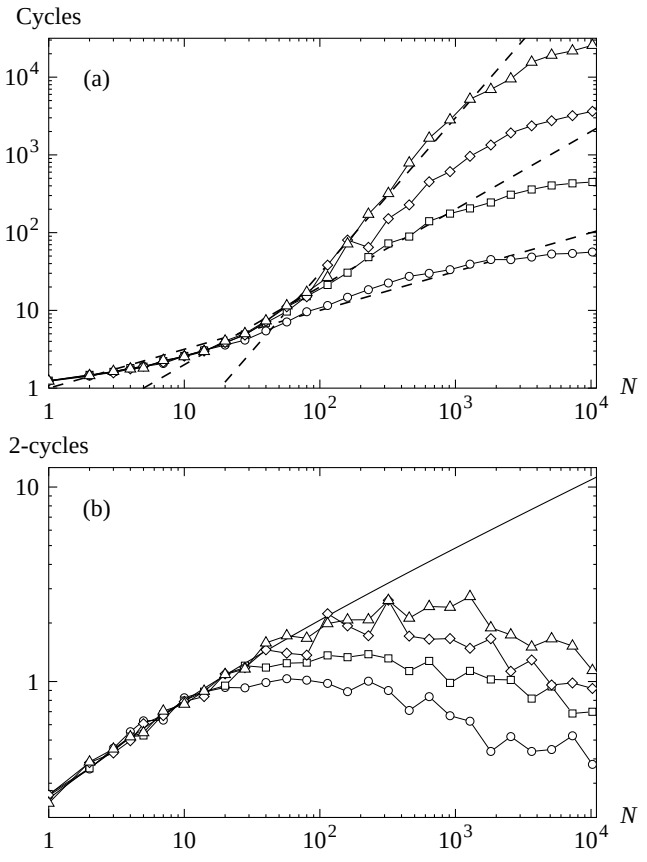


FIG. 2: The number of observed attractors (a) and 2-cycles (b) per network as functions of $N$ for different numbers of trajectories $\tau$: 100(circles), $10^3$(squares), $10^4$(diamonds), and $10^5$(triangles). In (a), the dashed lines have slopes 0.5, 1, and 2. In (b) the solid line shows $\langle C_2 \rangle_N$, the true number of 2-cycles. In the high end of (b), $1\sigma$ corresponds to roughly 20%.

behavior with 1000 trajectories. That the true growth is faster than linear has now been firmly established [12].

To closer examine the problem of biased undersampling, we have implemented the network reduction algorithm from [11], and gathered statistics on networks with $N \lesssim 10^4$. We repeated the simulations for different numbers of trajectories $\tau$, with $100 \leq \tau \leq 10^5$. For each $N$ and $\tau$, $10^3$ network realizations were examined, and we discarded configurations if no cycle was found within $2^{13}$ time steps. The results are summarized in Fig. 2a.

For $\tau = 100$, the number of attractors follows $\sqrt{N}$ remarkably well, considering that $\tau = 10^3$ gives the quite different $N$ behavior seen in [11]. If we extrapolate wildly from a log–loglog plot, $e^{0.3\sqrt{N}}$ fits the data rather well.

As another example of how severe the biased undersampling is, we have included a plot of the number of 2-cycles found in the simulations (Fig. 2b). The underlying distribution is less uniform than for the total number of attractors, so the errors are larger, but not large enough to obscure the qualitative behavior. The number of observed 2-cycles is close to $\langle C_2 \rangle_N$ for low $N$, but as $N$ grows, a vast majority of them are overlooked. As expected, this problem sets in sooner for lower $\tau$, although the difference is not as marked as it is for the total number of attractors.

## SUMMARY

We have introduced a novel approach to analyzing attractors of random Boolean networks. By applying it to Kauffman networks, we have proven that the number of attractors in these grows faster than any power law with network size $N$. This is in sharp contrast with the previously cited $\sqrt{N}$ behavior, but in agreement with recent findings.

For the Kauffman model, we have derived an expression for the asymptotic growth of the number of $L$-cycles, $\langle C_L \rangle_N$. This expression is corroborated by statistics from network simulations. The simulations also demonstrate that biased undersampling of state space is a good explanation for the previously observed $\sqrt{N}$ behavior.

We wish to thank Carsten Peterson, Bo Söderberg, and Patrik Edén for valuable discussions. This work was in part supported by the National Research School in Genomics and Bioinformatics.

* bjorn@thep.lu.se
† carl@thep.lu.se

[1] S. Bornholdt and K. Sneppen, Phys. Rev. Lett. **81**, 236 (1998).
[2] N. Lemke, J. C. M. Mombach, and B. E. J. Bodmann, Physica A **301**, 589 (2001).
[3] A. Bhattacharjya and S. Liang, Phys. Rev. Lett. **77**, 1644 (1996).
[4] R. J. Bagley and L. Glass, J. Theor. Biol. **183**, 269 (1996).
[5] C. Oosawa and M. A. Savageua, Physica D **170**, 143 (2002).
[6] J. J. Fox and C. C. Hill, Chaos **11**, 809 (2001).
[7] M. Aldana, S. Coppersmith, and L. Kadanoff (2002), submitted to Springer Applied Mathematical Sciences Series, http://arXiv.org/abs/nlin.AO/0204062.
[8] S. A. Kauffman, J. Theor. Biol. **22**, 437 (1969).
[9] U. Bastolla and G. Parisi, J. Theor. Biol. **187**, 117 (1997).
[10] U. Bastolla and G. Parisi, Physica D **115**, 219 (1998).
[11] S. Bilke and F. Sjunnesson, Phys. Rev. E **65**, 016129 (2001).
[12] J. E. S. Socolar and S. A. Kauffman (2002), submitted to Phys. Rev. Lett., http://arXiv.org/abs/cond-mat/0212306.
[13] B. Derrida and Y. Pomeau, Europhys. Lett. **1**, 45 (1986).
[14] B. Derrida and D. Stauffer, Europhys. Lett. **2**, 739 (1986).
[15] H. Flyvbjerg, J. Phys. A **21**, L955 (1988).