On the number of cycles and other combinatorial properties of random network models of gene regulation

Carlos Gershenson[1,2], Hyobin Kim[1], Octavio Zapata[1]
[1]Centro de Ciencias de la Complejidad , Universidad Nacional Autónoma de México, México
[2]Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas, Universidad Nacional Autónoma de México, México
cgg@unam.mx hyobin.kim@c3.unam.mx octavio.zapata@c3.unam.mx

Genes are the fundamental unit of inheritable information. A gene is a part of the genomic sequence that encodes how to produce (synthesise) either a protein or some RNA (gene products). Gene product synthesis is called gene expression. Not all genes are expressed at the same time. The expression of each gene is affected by the expression of other genes in a process called gene regulation. This gives rise to a network-like structure called genetic regulatory network.

Random Boolean networks were proposed by Kauffman in [] as models of genetic regulatory networks. The expression of each gene is represented using one bit: 1 represents the gene is expressed, and 0 the gene is not expressed. Let $\{0,1\}^n$ be the set of all binary words of length n. Each index $i \in I := \{1,\dots,n\}$ represents a gene. A point $x = (x_1,\dots,x_n) \in \{0,1\}^n$ is called a *state*, where x_i represents the expression of gene i.

A *Boolean network* with parameters n and k, $1 \leq k \leq n$, consists of a family $\mathbf{y} = \{y(i) : i \in I\}$ of subsets of genes $y(i) = \{i_1,\dots,i_k\} \subseteq I$, and a family $\mathbf{f} = \{f_i : i \in I\}$ of functions $f_i\colon \{0,1\}^k \to \{0,1\}$. The k genes in $y(i)$ are called the *regulators* of i. Notice that the number k of regulators is the same for all genes. A *random Boolean network* is a Boolean network $(\mathbf{y},\mathbf{f})$, where $y(i)$ and f_i are chosen randomly and independently of each other.

A *directed graph* G consists of a finite set of elements $V(G)$ called vertices, and a set of pairs of elements $E(G) \subseteq V(G)\times V(G)$ called directed edges. For any directed edge $(u,v) \in E(G)$, we say that u is a *predecessor* of v, and v is a *successor* of u. A *directed pseudoforests* is a disjoint union of directed graphs where each vertex has a unique successor. Every function $F\colon X \to X$, from a set X to itself, defines a directed pseudoforest with the elements of X as vertices, and $\{(x,F(x)) : x \in X\}$ as directed edges.

Similarly, every Boolean network $(\mathbf{y},\mathbf{f})$ defines a directed pseudoforests on the state-space $\{0,1\}^n$. More presicely, we define $F\colon \{0,1\}^n \to \{0,1\}^n$

$$F(x) := \big(f_1(x_{y(1)}),\dots,f_n(x_{y(n)})\big), \quad x \in \{0,1\}^n,$$

where $x_{y(i)} := (x_{i_1},\dots,x_{i_k})$ is the restriction of state x to the genes in $y(i) = \{i_1,\dots,i_k\}$. Thus, we identify a Boolean network $(\mathbf{y},\mathbf{f})$ with the mapping F and a directed pseudoforest on $\{0,1\}^n$.

Now we turn to more general gene regulation models, where the expression of each gene is represented by an element in $\{0,1,\dots,q-1\}$.

A *fuzzy network* with *base* $q \geq 2$ consists of a pair $(\mathbf{y},\mathbf{f})$, $\mathbf{y} = \{y(i) \subseteq I : i \in I\}$, $\mathbf{f} = \{f_i : i \in I\}$, with $f_i\colon \{0,1,\dots,q-1\}^k \to \{0,1,\dots,q-1\}$, such that

$$\begin{aligned} f_i(x_{i_1},\dots,x_{i_k}) =& \big(x_{i_1} \wedge f_i(x_{i_1},\dots,x_{i_k})\big) \\ &\vee \big(\neg x_{i_1} \wedge f_i(\neg x_{i_1},x_{i_2},\dots,x_{i_k})\big), \end{aligned} \tag{1}$$

where $x \wedge x' := \min\{x,x'\}$, $x \vee x' := \max\{x,x'\}$, and $\neg x := q-1-x$. Every fuzzy network with base $q=2$ is precisely a Boolean network. Fuzzy networks have been recently used to study some aspects related to the differentiation of cells from the immune system [], and also to explain the importance of certain gene products associated with the development of metabolic syndrome and type 2 diabetes [].

A *random fuzzy network* is a fuzzy network $(\mathbf{y},\mathbf{f})$, where $y(i)$ and f_i are chosen randomly and independently of each other. Random fuzzy networks are also known as random multiple-valued networks (see Dubrova et al. []), or as random multistate networks (see Wittmann et al. []). They are a particular class of random networks with multiple states (see Solé et. al. []).

Two directed graphs G and H are *isomorphic* if there is a function $\psi\colon V(G) \to V(H)$, such that for all $u,v \in V(G)$, $(u,v) \in E(G)$ if and only if $(\psi(u),\psi(v)) \in E(H)$. Fuzzy networks with base q are directed pseudoforests on $\{0,1,\dots,q-1\}^n$. Two fuzzy networks are said to be *equivalent* if they are isomorphic as directed pseudoforests. Let $D(n,q)$ be the number of non-equivalent random fuzzy networks with $k=n$. By a result of Bach et al. [], for all $q \geq 2$, we have

$$\sqrt{n} \ll \log D(n,q) \ll \frac{n}{\log\log n}, \quad \text{as} \quad n \to \infty,$$

where $A \ll B$ denotes $|A| \leq cB$ for some constant c.

A *walk* in a directed graph G is a sequence of vertices $v_1,v_2,\dots \in V(G)$, so that for all $j \geq 1$, $(v_j,v_{j+1}) \in E(G)$.

A walk where all vertices are distinct is called a *path*. Paths are necessarily finite walks. A finite walk where the first and the last vertex are the same is called a *cycle*. The number of distinct vertices in a cycle is the called the *length* of the cycle. The average number and length of cycles of states are two well-studied combinatorial parameters for random Boolean networks. Let $C_k(n,q)$ and $L_k(n,q)$ denote respectively the average number and length of cycles on a random fuzzy network with n genes and base q, where the number k of regulators per gene is fixed to some constant value. We write $C(n,q)$ and $L(n,q)$ in the extremal case, when $k=n$.

By a result of Kruskal [], for all $q \geq 2$, we have

$$C(n,q) = \frac{1}{2}\log q^n + \left(\frac{\log 2 + C}{2}\right) + o(1), \quad \text{as} \quad n \to \infty,$$

where $C = 0.5772\ldots$ is Euler-Mascheroni constant. In the context of gene regulation, this formula is well-known for the Boolean case $q = 2$, but seems to be new for random fuzzy networks with base $q > 2$. For all $q \leq 5$, we have

$$\sqrt{n} \ll C(n,q) \ll n, \qquad \text{as} \qquad n \to \infty.$$

By a result of Flajolet et al. [], for all $q \geq 2$, we have $L(n,q) = \sqrt{\pi q^n/8} + o(1)$, as $n \to \infty$. This is again well-known in the literature of random Boolean networks, but new for random fuzzy networks with base $q > 2$.

Families/Brownian bridge asymptotics (see Aldous et al. []) ... number of all families equals number of Łukasiewicz bridges.... Formula in book Analytic combinatorics....